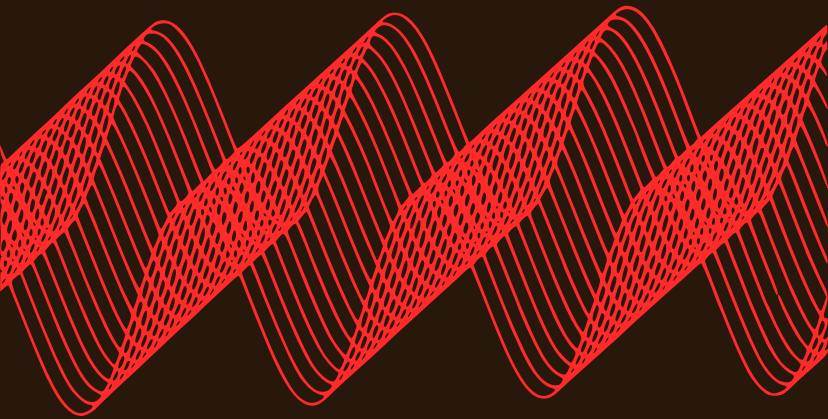
LASER AGE IN OPTICS

L.V. Tarasov

Mir Publishers
Moscow




ИЗДАТЕЛЬСТВО
МИР

Л. В. Тарасов

Оптика, рожденная лазером

Издательство «Просвещение»
Москва

L. V Tarasov

Laser Age in Optics

Translated
from the Russian
by V. Kisin, Cand. Sc. (Phys.)

Mir Publishers
Moscow

First published 1981
Revised from the 1977 Russian edition

На английском языке

Contents

Nonlinear optics

Historical background

Preface

There can be no doubt that the laser represents one of the most remarkable scientific and technical milestones of the 20th century. The dramatic growth of laser technology began in 1960, when the first successful laser was reported. Lasers are now used in the most diverse fields: biology and medicine, cybernetics and computer circuitry, communications and radar systems, industrial processes, and measurements of various types.

A laser is a very special light source, greatly different from incandescent lamps, fluorescent lights, and so forth. In contrast to other sources of light, the laser's radiation is characterized by a high degree of ordering of the light field. This idea is expressed by saying that it has a high degree of coherence. A laser can be treated as a sort of an "optical radio transmitter"; in comparison, all other light sources generate only "optical noise"

Until the advent of the laser, the radiofrequency range and the optical range differed greatly with respect to coherence: radiophysics widely used coherent waves while optics had only incoherent light at its disposal. A textbook was the only place where light

could "exist" as sine waves. Such light waves became real only when the laser was invented.

Laser optics is the optics of coherent radiation. Although this new branch appeared only twenty years ago, it has already given rise to a number of surprises: novel and very unusual optical phenomena were found and then put to use in extremely interesting applications.

The two newest directions of coherent optics are especially significant: *optical holography* and *nonlinear optics*. These two make up the subject of this book.

From Incoherent to Coherent Optics

1. Waves and Their Interference

It should be helpful for our purposes to precede the discussion of optics by considering some very general properties of waves and their interference. The physical nature of waves can be ignored for the moment.

Monochromatic plane wave. A monochromatic plane wave propagating without attenuation is the simplest type of wave. It can be described by the formula

$$\rho(x, t) = A \cos\left[2\pi\nu\left(t - \frac{x}{v}\right)\right] \tag{1.1}$$

The direction of x is assumed positive in the direction of propagation of the wave. In the case of elastic waves, $\rho(x, t)$ is the displacement of particles of the medium at the point x from the equilibrium position, at each moment t; ν is the wave frequency; A is the amplitude; and v is the velocity of wave propagation.

Equation (1.1) can be derived in a very simple manner. Let the source of harmonic oscillations be located at the origin of the reference frame (at point $x = 0$), where the displacement at moment t is $\rho_0(t) = A \cos(2\pi\nu t)$. Consider now a point at a distance

x from the origin. A wave takes the time x/v to cover this distance. Consequently, the displacement $\rho(x, t)$ at point x and at time t must be identical to that at point 0 and at time $t - (x/v)$. Hence, $\rho(x, t) =$

$$= \rho_0\left(t - \frac{x}{v}\right) = A\cos\left[2\pi\nu\left(t - \frac{x}{v}\right)\right].$$

By introducing the wavelength $\lambda = v/\nu$, one rewrites Eq. (1.1) in the form

$$\rho(x, t) = A\cos\left[2\pi\left(\nu t - \frac{x}{\lambda}\right)\right] \tag{1.2}$$

Equations (1.1) and (1.2) describe a monochromatic plane undamped wave.

This wave is said to be *monochromatic* because it involves a single frequency ν. Furthermore, it is *plane* since the displacement ρ is a function of a single spatial coordinate (namely, x); the wave front (i.e. the locus of points all of which have the same phase at any moment) is a plane perpendicular to the axis x. And finally, the wave in question is *undamped* since its amplitude A is constant at all points along the propagation direction. The energy of oscillation is known to be proportional to the square of the amplitude. Consequently, the constancy of the amplitude along the propagation direction signifies that the wave energy is transferred from one point to the next without losses, that is, no damping takes place.

Further in the text we use a quantity called *intensity*, denoted by I and defined as the amount of energy transported by a wave per unit time across a unit area oriented normally to the propagation direction at the point of observation. In complete

analogy to the energy of oscillations the intensity of a wave is proportional to its amplitude squared:

$$I \sim A^2 \tag{1.3}$$

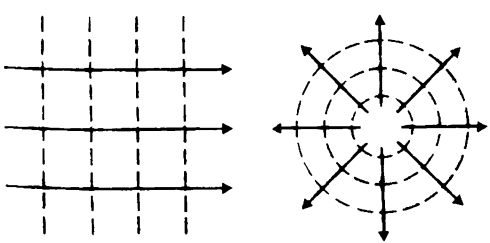
Monochromatic spherical wave. A point source placed in an isotropic medium (a medium whose properties are independent of direction) generates a wave with a spherical front, referred to as a spherical wave. The difference between plane and spherical waves is shown in Fig. 1 (arrows indicate the directions of propagation, and dashed lines trace wavefront cross-sections). Note that only the plane wave can be said to have a definite direction of propagation.

Under the assumption that a spherical wave is monochromatic and its energy is not absorbed by the medium, the corresponding wave equation is

$$\rho(r, t) = \frac{a}{r} \cos\left[2\pi\left(\nu t - \frac{r}{\lambda}\right)\right] \tag{1.4}$$

Here r is the distance between the source and the observation point, also termed the radius of the

Fig. 1



spherical wave front; a is the oscillation amplitude close to the source; and a/r is the wave amplitude at a distance r from the source. As r increases, the area of the wave front (i.e. of the sphere) rises proportionally to r^2, and the wave intensity consequently diminishes as $1/r^2$, since the total energy transported by the wave per unit time across the whole sphere does not depend, in a nonabsorbing medium, on the sphere radius. Hence, the wave amplitude must be proportional to $1/r$. Note that wave attenuation in the above case is not caused by absorption in the medium but only by wavefront divergence. A plane wave has zero divergence, and for a spherical wave the divergence is of maximum value.

Real waves. Rigorously speaking, oscillations generated by real sources of waves are never harmonic, so that real waves are never monochromatic. A real wave front may have a very complicated shape, far from planar or spherical. In addition, the shape may change in the course of wave propagation. Finally, it should be kept in mind that to a certain extent all real mediums are absorbing, so that the attenuation of a real wave is related not only to its divergence but also to its absorption in the medium.

It is essential, however, that any real wave can be represented mathematically as a sum of a set of distinct plane or spherical waves.

Any nonharmonic oscillation is representable as a sum of harmonic vibrations with different frequencies and amplitudes. Let us assume that a real oscillation is given by a sum of harmonic oscillations with frequencies in the interval $\Delta\nu$ around a mean frequency

ν_0. The ratio $\Delta\nu/\nu_0$ characterizes the *degree of nonmonochromaticity* of the real wave in question. A wave is termed quasimonochromatic, that is, nearly monochromatic, if $\Delta\nu/\nu_0 \ll 1$.

In what follows we consider a real wave, in the sense outlined above, as a sum of monochromatic plane waves. An individual monochromatic plane wave is characterized by a definite frequency and a definite direction of propagation (no divergence). A real wave may be only partially monochromatic and to a certain degree divergent. We have already mentioned the degree of nonmonochromaticity as a characteristic of a real wave. Likewise, the *degree of divergence* of a wave can be introduced. These characteristics are discussed later in the text for light waves.

Interference of waves. Interference is known to take place when two (or more) waves are superposed. In order to find out the essential features of this phenomenon, consider the simplest case of two monochromatic plane waves with equal frequencies. Let us analyze two situations.

The first example is the superposition of a wave described by Eq. (1.2) and a similar one propagating in the opposite direction. The second wave is described by the same equation (1.2) but with the minus sign in front of x/λ replaced by a plus. Mathematically, this superposition takes the form

$$p(x, t) = A\cos\left[2\pi\left(\nu t - \frac{x}{\lambda}\right)\right] + A\cos\left[2\pi\left(\nu t + \frac{x}{\lambda}\right)\right]$$

$$p(x, t) = 2A\cos(2\pi\nu t)\cos\frac{2\pi x}{\lambda} \qquad (1.5)$$

This is an equation of the so-called *standing wave.* It should be emphasized that some points on a standing wave are at rest. These are the points for which $\cos(2\pi x/\lambda) = 0$; they are called the *nodes* of a standing wave, and their coordinates are

$$x = \frac{\lambda}{2}\left(\frac{1}{2} \pm n\right)$$

where n are integers.

At the same time there are points (located halfway between each pair of neighbouring nodes) whose amplitude of oscillation is twice that of each of the waves; these are called the *antinodes* of the wave.

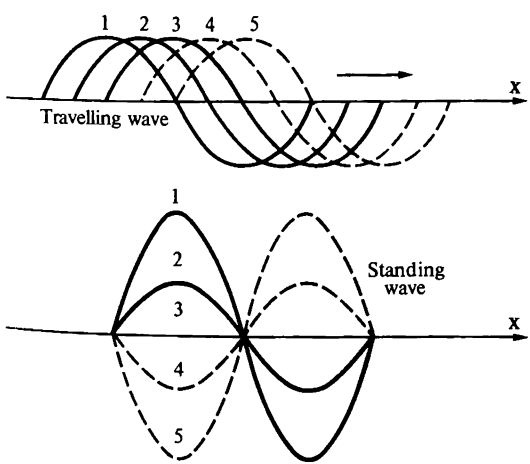
There is definitely much to be surprised about in this result: we have found that the interference (summation) of two travelling waves propagating in opposite directions results in the energy of these waves not being transferred at all to some fixed points of the medium, while to other points it is transferred in apparently excessive amounts (indeed, an amplitude of $2A$ at the antinodes implies that the energy of oscillation at these points is $4A^2$, and therefore is twice the total energy of oscillation of points close to the sources of both waves).

A standing wave is the simplest example of wave interference. The difference between a standing and an ordinary (travelling) wave is illustrated in Fig. 2, where the displacement ρ is plotted as a function of coordinate x for five different moments of time.

Consider now the case of superposition of two monochromatic plane waves propagating at an angle α. The waves' frequencies and amplitudes are identical (Fig. 3).

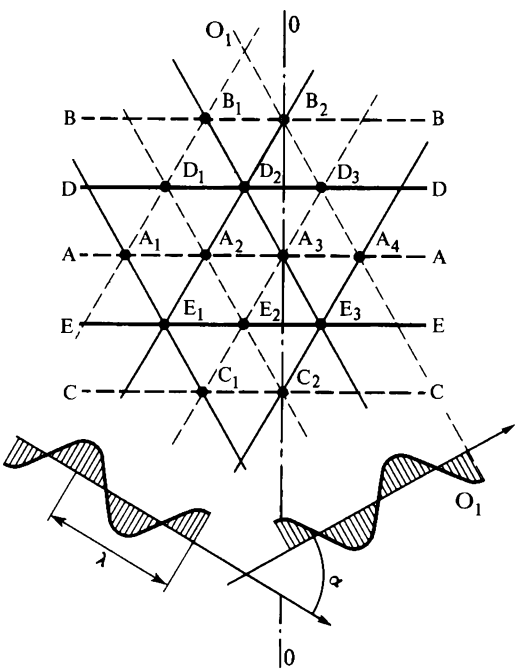
Thin straight lines in Fig. 3*a* trace the system of wave fronts of the interfering waves considered independently. Heavy solid lines trace the fronts on which the phase of the wave is equal to $2\pi n$ (n are integers) so that the cosine in Eq. (1.2) is unity, while dashed lines are fronts with phase $\pi(2n + 1)$ so that the cosine is minus unity. The two indicated wave fronts intersect either with identical phase (points D_1, D_2, D_3, E_1, E_2, E_3 oscillating with amplitude $2A$), or exactly out-of-phase (points A_1, A_2, A_3, A_4, B_1, B_2, C_1, C_2 in which amplitude is zero). As the waves propagate, points D_1, D_2, D_3 move along the line DD, while points E_1, E_2, E_3 move along EE. Correspondingly, other points trace the lines AA, BB, CC.

Fig. 2

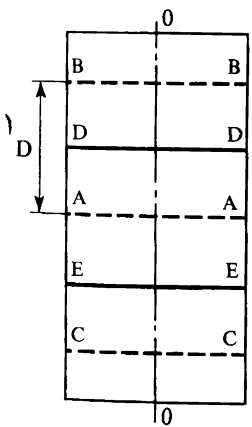


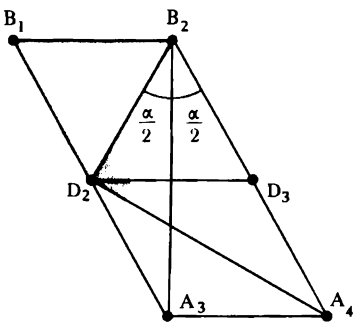
As a result of interference, therefore, stationary regions are formed where points are at rest, and regions are formed where points oscillate at twice the amplitude. These regions are traced in Fig. 3*a* by dashed and solid heavy lines, respectively (in reality, these straight lines are sections of planes of constant amplitude).

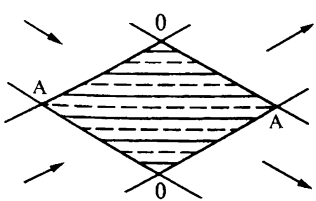
Fig. 3



Let us cut the interference region by a plane perpendicular to the plane of the drawing. The trace of this plane is shown in Fig. 3*a* as line *OO*; the plane "faces" the reader in Fig. 3*b* (dashed lines indicate the regions with zero intensity of the resultant wave, and solid lines show regions of maximum intensity). The distance between neighbouring lines of zero intensity is







denoted by D; obviously, the distance between the maximum intensity lines is the same.

In order to determine D, consider Fig. 3*c*, which is a fragment of Fig. 3*a*. It is easily seen that $D_2A_4 = \lambda$, angle $B_2D_2A_4$ is 90° and angles $D_2B_2A_3$ and $A_3B_2A_4$ are each equal to $\alpha/2$. Hence, $B_2A_4 = D_2A_4/\sin\alpha = \lambda/\sin\alpha$. Moreover,

$$B_2A_3 = B_2A_4 \cos\frac{\alpha}{2} = \frac{\lambda \cos\frac{\alpha}{2}}{\sin\alpha} = \frac{\lambda}{2\sin\frac{\alpha}{2}}$$

Finally we obtain

$$D = B_2A_3 = \frac{\lambda}{2\sin\frac{\alpha}{2}} \tag{1.6}$$

Note that for $\alpha = \pi$ we have $D = \lambda/2$. This is the distance between the neighbouring nodes of the standing wave.

An interference pattern is often observed in a plane perpendicular to the direction of propagation of one of the plane waves. Such a plane is, for example, plane O_1O_1 in Fig. 3*a*. Obviously, the distance d between the lines of zero intensity in this plane is the segment B_2A_4 (see Fig. 3*c*). Therefore,

$$d = B_2A_4 = \frac{\lambda}{\sin\alpha} \tag{1.7}$$

The general arrangement necessary to obtain an interference pattern is shown in Fig. 3*d*. Two wave

trains (their boundaries are shown arbitrarily) interfere within a certain volume shown in the figure by rhomb *AOAO*. Solid horizontal lines trace regions of maximum intensity and dashed lines trace zero-intensity regions. The figure demonstrates that interference results in a spatial redistribution of energy. Outside of rhomb *AOAO* the energy of oscillations is uniformly distributed within each beam of waves, while the distribution within the interference region is essentially nonuniform, with energy concentrating in the vicinity of solid horizontal lines in the figure.

Interference: Summary. The examples discussed above demonstrate that, *first of all*, superposition of waves results in a spatial redistribution of the oscillation energy; in other words, the intensity of waves is redistributed in space. Regions with zero intensity and those with intensity above the intensity of overlapping waves are produced. *Second*, this spatial redistribution of energy is found to be time-independent. This means that a stable pattern of fixed interference fringes is formed.

The phenomenon of wave interference reduces therefore to a spatial redistribution of intensities of waves, which results in formation of a fixed pattern of interference fringes.

It must be noted in conclusion that interference of two plane waves constitutes one of the elementary cases of interference. It is also possible to analyze the cases of interfering plane and spherical waves, two interfering spherical waves, or waves with still more complex wavefronts. Obviously, the more complicated the wavefronts, the more complex the interference fringe patterns are.

2. Is Interference an Inevitable Result of Wave Superposition?

Superposition of two monochromatic plane waves always produces an interference pattern. Is a similar statement correct for any two arbitrary real waves? Is interference an inevitable corollary of superposition of real waves?

The answer is no. There are numerous physical phenomena in which superposition of waves produces no interference. This is the field of optics, i.e. of *light waves*. Our day-to-day experience indicates that as a rule no interference is observed when light waves superpose. The resultant intensity of several superposed light waves is simply the sum of intensities of the component waves. (Before the advent of the laser, very special conditions were required to observe interference patterns, such as thin film colour fringes or Newton's rings.) The reason lies in specific features of the range of wavelengths that are used in optics

The spectrum of electromagnetic waves. We remind the reader that the spectrum of electromagnetic waves (also referred to as the radiation spectrum) is separated into two main parts: waves emitted by electric oscillators ($\lambda > 10^2$ μm) and those emitted by molecules, atoms, and nuclei ($\lambda < 10^2$ μm). The *first part* includes, among other types of radiation, radio waves and microwave radiation, which lies directly next to the boundary wavelength $\lambda = 10^2$ μm. The *second part* of the spectrum begins with the optical range, which covers wavelengths from 10^2 to 10^{-2} μm. The range of wavelengths below 10^{-2} μm corresponds to

the X-ray range, followed at shorter wavelengths by the range of γ-radiation. The optical range is represented by radiation emitted by molecules and atoms, and the X-ray range by that of atoms and nuclei.

The optical range is subdivided in its turn into three parts: visible radiation ($\lambda = 0.75$-0.4 μm), infrared radiation ($\lambda > 0.75$ μm), and ultraviolet radiation ($\lambda < 0.4$ μm). The infrared (IR) part of the spectrum is mostly the radiation of molecules, while the visible and ultraviolet (UV) parts are represented by emission from atoms.

One important feature must be emphasized in this context: the main distinction between the electromagnetic waves of the optical and short-wave ranges and those of the radio and microwave ranges lies in the process of wave generation. In the case of radio and microwave frequencies the process of generation is based on regularly repeated motions of electrons in oscillatory systems, while generation in the optical range is produced by quantum transitions in individual molecules and atoms.

A brief reminder concerning *electromagnetic waves* is in order here.

Two vectors oscillate in an electromagnetic wave: the electric and magnetic vectors $\vec{E}$ and $\vec{B}$, respectively, The first of these is called the vector of electric field strength, and the second the magnetic induction vector. The vectors are mutually perpendicular, and normal to the direction of wave propagation (electromagnetic waves are transverse). The wave propagation velocity in a medium is

$$v = \frac{c}{n} \tag{2.1}$$

where n is the absolute refractive index of the medium, and c is the velocity of light in a vacuum ($c = 3 \cdot 10^8$ m/s).

The vectors $\vec{E}$ and $\vec{B}$ are equivalent components of an electromagnetic wave, but the photochemical, photoelectric, and physiological effects of light are mostly produced by the electric vector $\vec{E}$. For this reason we restrict the following discussion to $\vec{E}$.

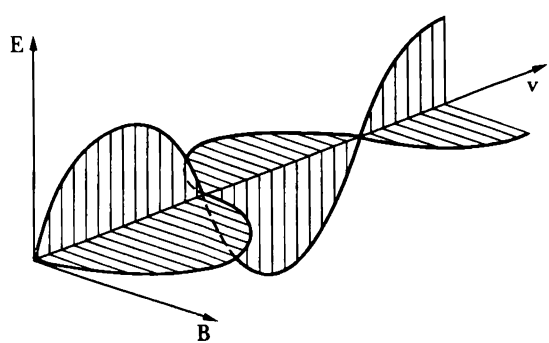
Figure 4 gives a clear representation of a plane electromagnetic wave. It shows that the vectors $\vec{E}$ and $\vec{B}$ oscillate in different planes and are in phase. One of these planes is chosen as the *plane of polarization* of the wave. The electric vector being more important for applications, we choose to consider the $\vec{E}$-plane as the polarization plane.

How light is generated. Processes leading to emission of light are numerous and varied. For instance, light is emitted by a decelerated charged particle, such as an electron, as a result of interaction with atomic and interatomic fields (so-called bremsstrahlung), by an electron moving in a medium at a velocity above the velocity of propagation of electromagnetic waves in this medium (Vavilov-Cherenkov radiation), and in electron-positron collisions (annihilation emission). The most typical process of light generation, however, is that of transition from excited states to non-excited (or rather, less excited) states in atoms (or molecules) of the emitting substance. This mechanism is operative when light is emitted by a match flame, an incandescent lamp, a fluorescent light, and finally a laser. For this reason, our attention will be concentrated exclusively on this mechanism of light generation.

Whatever the mechanism, light is always emitted in the form of very specific particles, so-called *photons.* This becomes obvious when the process of light emission is treated as a result of atomic (or molecular) transitions from excited to non-excited (ground) states: indeed, a transition in an individual atom or molecule generates an individual light particle, that is, a photon.

The possible values of energy of an atom (or molecule) are discrete, which fact allows us to refer to a system of atomic (or molecular) energy levels. Let us assume that an atom undergoes a transition from a state with energy E_2 to a state with lower energy E_1 (from the E_2 energy level to the E_1 level). This transition generates a photon with energy $\varepsilon = E_2 - E_1$. It is readily apparent that a reverse transition must result in the elimination (absorption) of a photon with energy $E_2 - E_1$. The set of energies of photons that can be generated (or absorbed) by an atom is easily

Fig. 4



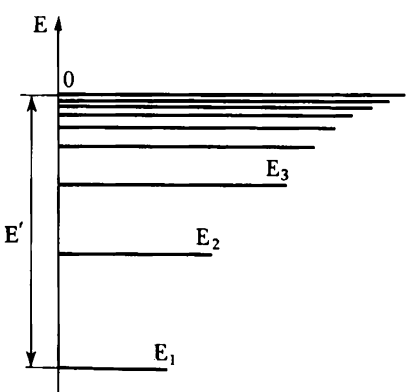
predicted once the system of energy levels of the atom in question is known.

Consider as an example the hydrogen atom, making use of the model that, as the reader is well aware, was suggested by the great Danish physicist Niels Bohr. According to this model, the energy levels in a hydrogen atom are given by the formula

$$E_n = -\frac{4\pi^2 me^4}{h^2 n^2} \tag{2.2}$$

where m is the electron mass, e is its charge, n is an integer, and h is a universal physical constant termed Planck's constant ($h = 6.6 \cdot 10^{-34}$ J·s). Energy levels in the hydrogen atom are shown in Fig. 5. Usually they are referred to as the electron energy levels in the hydrogen atom. Level E_1 corresponds to the ground

Fig. 5



state of the atom, in which an electron circles the nucleus in an orbit of minimum radius. Levels E_2, E_3, and so on, denote excited states of the atom. The zero point of the energy axis (the origin of the reference frame) corresponds to the ionized state in which the electron is outside of the atom. The energy of ionization of a hydrogen atom is given by

$$E' = \frac{4\pi^2 me^4}{h^2} = 13.55 \text{ eV}$$

(1 eV = $1.6 \cdot 10^{-19}$ J). Equation (2.2) indicates that the energies of photons that can be emitted by a hydrogen atom are found from the expression

$$\varepsilon = \frac{4\pi^2 me^4}{h^2}\left(\frac{1}{n^2} - \frac{1}{k^2}\right) \tag{2.3}$$

where n and k are integers and $n < k$. Assuming $n = 1$ and varying k, we obtain the possible photon energies produced by transitions to the ground state, that is, to the level E_1. Assuming $n = 2$, we obtain the spectrum of photons emitted in transitions to the level E_2. Assuming $n = 3$, we arrive at the spectrum for transitions to the level E_3, and so on.

Photons. Microparticles, and photons among them, are very peculiar physical objects; their behaviour is described not by classical mechanics but by so-called *quantum mechanics.* Many of the customary, age-old concepts formed as a result of observation and investigation of ordinary (classical) objects have to be discarded when microparticles are analyzed. For example, many questions lose their physical meaning,

such as: What does a photon look like? What are its components? What is its size?

We avoid the above "head-on" questions by choosing a bypass: what should be known about a photon in order for it to be "fixed" (described)?

First, the photon's energy ε, and *second*, its direction of propagation must be determined. This is equivaltnt to determining the photon's momentum $\vec{p}$ The momentum of a photon coincides with its direction of propagation, and the magnitude of the momentum is related to energy by the formula

$$p = \frac{\varepsilon}{c} \tag{2.4}$$

where c stands for the velocity of light in the medium.

And *finally*, the photon's polarization must be determined. This characteristic is quite similar to the corresponding one of a light wave. For instance, one can speak of a photon polarized in a specific plane. Two independent states of polarization of a light wave are known (we mean the states of wave polarization in two mutually perpendicular planes). Correspondingly, we refer to two states of polarization of a photon and characterize them by a parameter σ that assumes the value 1 for one state and 2 for the other one.

How can one picture a photon's polarization? In a word, what is it? We are used to constructing a clear, demonstrative image, a model of a phenomenon. However, it is very often impossible (in principle!) to work out descriptive images in the world of microparticles. In particular, it is impossible to describe the difference in the "appearance" of $\sigma = 1$ and $\sigma = 2$ photons. What, then, is the meaning of

polarization with respect to photons? The only legitimate answer is the following: if a photon is "extracted" from a light wave with polarization $\sigma = 1$, its polarization will likewise be $\sigma = 1$; if, however, it is extracted from a light wave with polarization $\sigma = 2$, its polarization is $\sigma = 2$. Nothing more specific (more descriptive) can be read into the concept of photon polarization.

Consequently, for a photon to be described, four quantities must be determined: three projections of momentum (p_x, p_y, p_z) and σ, which defines the photon's polarization. As can be seen from Eq. (2.4), this immediately gives the photon's energy:

$$\varepsilon = pc = c\sqrt{p_x^2 + p_y^2 + p_z^2}$$

Several photons are said to be in the same state if they have identical sets of the four characteristics p_x, p_y, p_z, σ. This set can be regarded, therefore, as a characterization of a *photon state*. A photon state changes if at least one of the four characteristics changes.

It is important that the characteristics of a photon state correspond to those of a monochromatic plane wave. Thus, a photon's momentum is directed along the wave propagation direction, and its polarization is that of the wave. As for the energy of the photon, it is given by the wave's frequency:

$$\varepsilon = h\nu \tag{2.5}$$

It can be said that a monochromatic plane wave is an ensemble of photons in the same state. Different

photon states correspond to different monochromatic plane waves.

Substitution of Eq. (2.5) into (2.4) yields

$$p = \frac{h\nu}{c} = \frac{h}{\lambda} \tag{2.6}$$

Equations (2.5) and (2.6) point to the particle-wave duality of a photon's property: they relate the *corpuscular* (ε, p) and *wave* (ν, λ) characteristics of a photon.

Fermions and bosons. The number of microparticles known nowadays is quite impressive. It includes about two dozen elementary particles (photon, neutrino, electron, proton, neutron, etc.), about the same number of antiparticles, as well as various atoms and atomic nuclei. It is noteworthy that all microparticles in nature are clearly classified into two groups by the character of their behaviour in an ensemble of similar particles (a physicist's way of saying the same is: "by their statistical properties"). Figuratively speaking, particles of the first group are extremely individualistic: once a state is occupied by a particle, no other particle of the same type can occupy the same state. In other words, these particles are governed by the one-state-one-particle rule. Microparticles of the second group behave quite differently: in an ensemble the number of particles per state is not only unlimited, but the greater the number of particles already occupying a state is, the higher the probability of finding a particle in this state.

Particles of the first group are called *fermions* (in honour of the Italian physicist Enrico Fermi), and

those of the second group—*bosons* (in honour of the Indian physicist Bose). A particle is either a fermion or a boson. This fundamental fact finds certain explanations in physics, but they are outside the scope of this book. A feature of extreme importance here is that photons are bosons. It will be demonstrated below that this fact (photons being governed by the Bose statistics) plays the most important role in optical phenomena.

Note that electrons are fermions, which explains the well-known fact that each energy level of an atom cannot contain more than a specific number of electrons (2, 8, 18, ...). The reason is: each atomic level corresponds to a specific number of electron states (2, 8, 18, ...). If electrons were not kept from occupying the same state in numbers above unity, all electrons would certainly be found on the lowest energy level; as a result, the existing variety of chemical elements would disappear. And this would mean the end of our world!

Photons and light waves. How can we reconcile the description of optical radiation in terms of photons with the existence of light waves? How can we "pass" from photons to light waves?

We have mentioned above that relationships (2.5) and (2.6) reflect the particle-wave duality of photons. In other words, a photon, as any other microparticle, possesses both corpuscular and wave properties. The wave properties of photons, however, are not sufficient to explain the existence of light waves. There is yet another reason, and a very important one: the behaviour of photons in an ensemble, in other words, the Bose statistics of photons.

Owing to these statistics, photons can populate any state in unlimited numbers. Moreover, the higher the population density of a state is, the higher the probability of occupying the state.

Let us mentally select a photon $p\sigma$-state, that is, a state characterized by a certain momentum $\vec{p}$ and polarization σ. This state corresponds to a definite monochromatic plane light wave (referred to, hereafter, as the $p\sigma$-wave) with frequency $\nu = \varepsilon/h = pc/h$. Let $N_{p\sigma}$ denote the number of photons per unit volume in a $p\sigma$-state.

If

$$N_{p\sigma} \gg 1 \tag{2.7}$$

that is, if the chosen photon state contains a very large number of photons, then the discontinuous nature ("granularity") of the radiation can be neglected and this radiation can be treated as a "continuous medium" (as a light wave). If, however, condition (2.7) is not satisfied, radiation must be considered as discrete, and the ensemble of photons cannot be treated as a light wave.

Recapitulating, a $p\sigma$ light wave exists if condition (2.7) holds; and if (2.7) is violated, the term "$p\sigma$ light wave" becomes meaningless. A high population density of photons in a $p\sigma$-state is the necessary and sufficient condition of the existence of a $p\sigma$ light wave.

It can be readily seen that classical photon waves can exist (i.e. light waves can exist) as a direct corollary of the Bose statistics of photons. Were this not the case, that is, were photons fermions, then not more than a single photon could occupy any one

photon state. Condition (2.7) would then be violated. It is worth mentioning that this is the reason why classical electron waves cannot be formed.

"*Disordered*" *light waves. Wave trains.* When generation of optical radiation is considered, different photon states have to be taken into account. Indeed, individual atoms (or molecules) of the emitting medium generate photons in a substantially independent manner; hence, the photons that appear differ in energy, momentum orientation, and polarization. Radiation "made up" of such different photons ("disordered radiation") cannot be treated as a monochromatic plane wave. A spread of photons over available states is characteristic of such "disordered" light waves.

A "disordered" light wave is often modelled by a set of so-called *wave trains.* Let us assume that the photons of which the radiation consists can be separated (mentally, of course) into groups each comprising a sufficiently large number of photons in identical states. The groups differ in the characteristics of photon states and in the degree of population of each state. Each such group is a wave train. In the simplest case, a wave train is illustrated by a "segment" of a monochromatic plane wave with characteristics corresponding to those of the photon state in question; the length of the train is determined by the population of the state (by the number of photons in the wave train): the greater this number, the greater is its similarity to a corresponding monochromatic plane wave. A wave train is schematically drawn in Fig. 6 (τ denotes the train duration, and τc — its spatial length).

Coherence of a light wave as a measure of its ability to interfere. The more "disordered" a light wave, the less capable it is of forming an interference pattern. In actual conditions, the ability of a wave to form interference fringes can be evaluated by measuring the contrast of the interference pattern, that is, the ratio

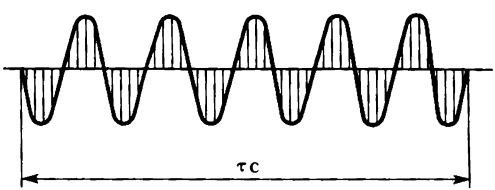
$$\eta = \frac{I_1 - I_2}{I_1 + I_2} \tag{2.8}$$

where I_1 is the intensity at the centre of a bright interference fringe, and I_2 is that at the centre of a dark fringe.

Contrast is at a maximum ($\eta = 1$) when $I_2 = 0$, and at a minimum ($\eta = 0$) when $I_1 = I_2$. In this case interference simply does not exist.

The higher the pattern contrast η (the closer it is to unity), the higher the wave's ability to form interference patterns is. We see, therefore, that numerous intermediate cases corresponding to different degrees of ability to interfere can be found between the limiting cases of ideal and destructive interference. These cases are illustrated in Fig. 7 (the ability to interfere diminishes from *a* to *c*).

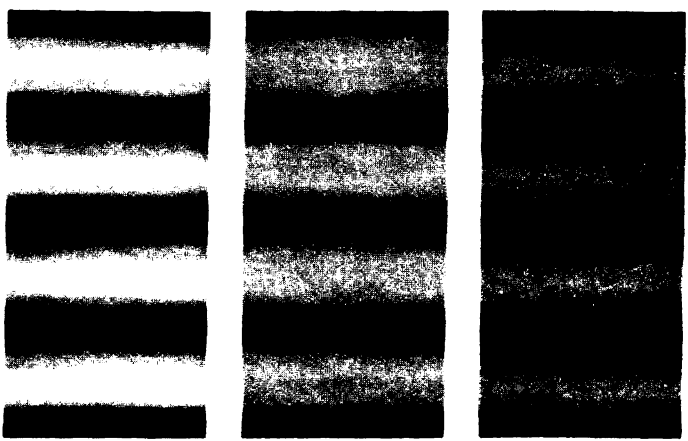
Fig. 6



An extremely important parameter, the *degree of coherence*, is introduced to characterize this ability of light to form interference patterns. The higher this degree, the greater its ability is. Correspondingly, a decrease in the degree of coherence corresponds to increased "disordering" of light.

Monochromatic plane waves are ideally coherent. Their ability to interfere is maximum, as is their "ordering" We may assume that the opposite case can also be realized, that is the case of ideally incoherent (absolutely "disordered") waves completely unable to form interference patterns. Obviously, actual situations correspond to various intermediate cases. Rigorously

Fig. 7



speaking, both ideally coherent and ideally incoherent waves are abstractions, and only partially coherent waves are a physical reality.

The degree of coherence of electromagnetic waves in the radio and microwave ranges is quite high because the process of generation of these waves is based on regularly repeated motions of electrons. Substantially lower coherence in the optical range is caused by the specific process of generation in this range (radiation of atoms and molecules). The optics existing before the advent of the laser is often called the *incoherent optics*; the term is definitely a very conventional one, since a small degree of coherence should have been mentioned. A rigorous approach would require that a higher degree of coherence be indicated for fluorescent lights than for incandescent lamps.

Lasers led to a revolution in optics. It was found that radiation by atoms and molecules can also be sufficiently coherent. The laser age in optics is the age of *coherent optics.*

At present, progress in coherent optics makes it possible to extend the methods of traditional radiophysics to the range of optics. The new branch of science that resulted from this extension is called *radiooptics.*

The degree of coherence of light waves and the character of photon populations. We want to clarify now what determines the degree of coherence of a light wave, that is, its ability to form interference fringes.

The decisive factor is the character of photon population of photon states. If coherence is ideal (the case of a monochromatic plane wave), all the photons

are in the same state, that is, have the same energy, the same direction of momentum, and the same polarization. On the contrary, ideal incoherence means sufficiently uniform distribution of photons over different states. The higher the *selectivity* (non-uniformity) in the photon population of the available states, or in other words, the greater the population density of some states at the expense of other states (which remain unoccupied or nearly unoccupied), the higher the radiation's coherence. In practical terms this means that the higher the degree of coherence of optical radiation is, the lower its nonmonochromaticity, the smaller its divergence, and the higher the degree of polarization.

The degree of nonmonochromaticity of a wave (see Sec. 1) is given by the ratio $\xi = \Delta\nu/\nu_0$, where ν_0 is the mean frequency and $\Delta\nu$ is the frequency range characterizing the "spread" in photon energies (if this "spread" is denoted by $\Delta\varepsilon$, then $\Delta\nu = \Delta\varepsilon/h$). The degree of divergence is characterized by the angle of the cone in which the wave propagates; this angle is referred to as the divergence angle. The smaller the angle of divergence, the closer the wavefront's shape is to planar. In practical cases the degree of polarization of light waves is found by means of *polarizers*, for instance by means of a special crystal that transmits only waves with a definite plane of oscillation of the electric field strength (fixed by the orientation of the crystal). The intensity of light transmitted by the polarizer is measured for different orientations of the crystal with respect to the beam propagation direction, and the minimum (I_2) and maximum (I_1) intensities are recorded. The degree of polarization is given then by

the ratio

$$\varkappa = \frac{I_1 - I_2}{I_1 + I_2} \tag{2.9}$$

Finally it should be noted that there exists a simple relationship between the degree of nonmonochromaticity of a light wave, ξ, and the time duration, τ, of wave trains:

$$\xi \sim \frac{1}{\tau \nu_0} \tag{2.10}$$

The quantity τ is generally referred to, in the special literature, as the *coherence time.* This is a very important characteristic of the degree of coherence of a light wave. Obviously, the longer the wave trains, the higher the degree of coherence; this is in perfect agreement with the dependence of coherence on the degree of selective population of photon states.

3. How Can We Generate Coherent Light Waves?

Coherent effects in pre-laser optics. It was mentioned above that any light wave is characterized by a certain degree of coherence. This is equally true for waves emitted by ordinary (non-laser) light sources, however small the degree of coherence of a light wave, in principle it can always be used to generate an interference pattern.

The truth of this can be demonstrated by using the concept of wave trains. Obviously, different wave trains do not give rise to interference when superposed; it is nevertheless possible to superpose parts of the

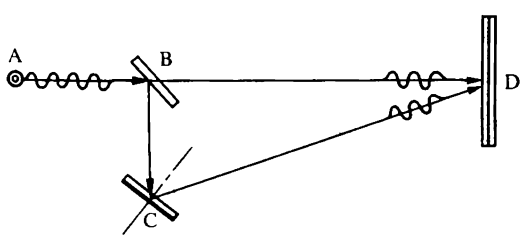
same train and thus obtain interference. The principal arrangement of an experiment making such interference possible is given in Fig. 8 (A is the source of light; B is a semitransparent mirror; C is a completely reflecting mirror; and D is a screen to observe the interference pattern). Mirror B splits the wave train, and mirror C realizes the superposition of the split parts of the train at point D. The following condition must be satisfied in order that parts of the same wave train superpose at point D:

$$L \ll \tau c \tag{3.1}$$

where τ is the wave train duration (coherence time), L is the difference between paths travelled by the parts of the wave train from the splitting point to the final point (in the case in question this is the difference between $|BC| + |CD|$ and $|BD|$). Interference is observed if condition (3.1) is satisfied.

Various methods are used in pre-laser optics to realize the principle illustrated in Fig. 8: Fresnel's biprism, Michelson's interferometer, observation of

Fig. 8



Newton's rings and thin film colour fringes, and so on. Interference fringes are observable in all these cases because L is sufficiently small (note that the train length in ordinary light sources is normally not more than one centimetre or even a fraction of a centimetre).

Photon population of states must be a controlled factor. The pertinent questions are: how can we generate a sufficiently coherent wave, and how can we increase the degree of coherence of a wave?

In principle, the answer is clear. It is necessary for the photons to occupy only selected states. This in turn requires that the photon population be a controlled characteristic.

Such control is indeed possible owing to the "boson nature" of photons, that is, to their tendency to populate predominantly just those states that already have a sufficiently high population density. Obviously, this property of photons can in principle be used to accumulate photons in specific states.

Let us discuss the methods by which the problem can be solved in actual conditions.

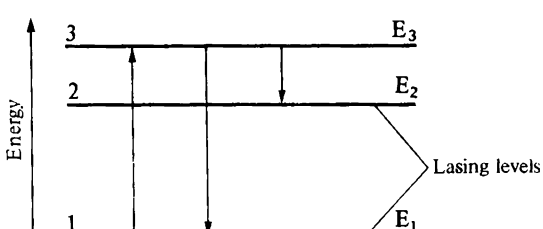
Generation of inverted population of the levels active centres. First of all, we want a material containing particles (atoms, ions, or molecules) with a special system of energy levels (also called terms). Such a material is called an *active medium*, and the particles are referred to as *active centres*. Figure 9 illustrates a characteristic system of levels of an active centre. It contains three levels (all other levels of the active centre play an insignificant role as far as the processes in question are concerned).

In ordinary conditions (in the initial state) all active centres have energy E_1, that is, are at level *1*. Assume now that somehow the active centres are excited so that a large enough number is raised to level *3*.

A microparticle cannot remain in an excited state for an indefinitely long time. Sooner or later, it inevitably undergoes a spontaneous transition to one of the less excited states; in the case in question, this means a transition to level *2* and then to level *1* (or directly from level *3* to level *1*). The released energy of excitation is then either transferred to other particles of the medium or emitted as a photon. It is impossible, in principle, to indicate the moment of time when an active centre spontaneously reduces its energy, but the probability of such a transition per unit time can be given. Note that this probability is independent of both the choice of the active centre and the time it has already spent in the excited state.

Let us denote the probability of spontaneous transition from level *3* to level *2* (the $3 \rightarrow 2$ transition) by $1/\tau_{32}$, and that of the transition $3 \rightarrow 1$ by $1/\tau_{31}$. By

Fig. 9



convention, constants τ_{32} and τ_{31} are called the lifetimes of the active centre on level *3* with respect to transitions $3 \rightarrow 2$ and $3 \rightarrow 1$. We also introduce the probability of transition $2 \rightarrow 1$ and denote it by $1/\tau_{21}$.

The system of energy levels of an active centre must be such that the condition

$$\tau_{32} < \tau_{31} \ll \tau_{21} \tag{3.2}$$

is satisfied. This means that active centres must leave level *3* fairly quickly and pass predominantly to level *2*, where they must be "trapped" for some time (as a rule, $\tau_{32} \approx 10^{-8}$s, and $\tau_{21} \approx 10^{-4}$-$10^{-2}$ s). If (3.2) holds, active centres on level *2* can be accumulated in numbers exceeding those on level *1*. It is said in this case that an *inverted* population of levels *1* and *2* is reached.

Stimulated (forced) emission of photons. Let us assume that a photon passes in the vicinity of active centres with inverse population of levels *1* and *2*, and that the photon's energy ε is equal to the difference in energies of these levels, $\varepsilon = E_2 - E_1$. What result can be expected?

First, the photon may be absorbed by one of those active centres that are still on level *1*, so that the number of active centres on level *2* increases by unity. *Second*, the photon in question may "stimulate" one of the active centres on level *2* to drop to level *1*; this generates one additional photon.

The physical nature of the second process is clear. It results in an increase in the number of photons in a certain photon state, whereby the photon emitted by an active centre is added to the initial photon. This is a direct consequence of the bosonic nature of photons,

that is, their tendency to accumulate in the same photon state. This fundamental property "makes" the photons "trigger" such transitions in the material, which produce new (secondary) photons. It must be emphasized that the secondary photon is in exactly the same state as the stimulating photon.

This process of photon emission differs in principle from spontaneous emission because here the transition is controlled by an external photon that "triggers" the emission. Consequently, one distinguishes between the *spontaneous* and *stimulated* (forced) emission.

The author once encountered a simile comparing stimulated emission with a process in which a pear falling off a tree shakes the branches and thereby forces down other pears in its wake. This comparison is justified, in a certain sense, since it demonstrates the difference between spontaneous and stimulated emission. The fall of the first pear happens as if "by itself" and we get a "spontaneous pear" The "stimulated pears", on the other hand, fall not by themselves but owing to a controlling factor: they are shaken off by other pears. In analogy to the case of photons, a "spontaneous pear" may serve as a trigger for the generation of "stimulated pears" This analogy should not, however, be taken too literally. For instance, it must be kept in mind that before being "triggered", a pear is an entity present on the branch, while a stimulated photon is generated in the very process of emission (it did not exist earlier). The initial photon does not actually "force another to follow suit", but triggers an appropriate transition in an active centre, which generates a stimulated photon.

The following interpretation can be given to the process of stimulated emission in the "language" of

trains. An initial train interacts with excited active centres and "triggers" transitions generating secondary trains. These secondary trains are generated essentially in phase with the primary one, which can be said to form a new, longer train. This lengthening of a train signifies an increase in the degree of coherence of the wave, requiring at the same time that the primary and secondary trains be in phase.

It must be emphasized that the generation of a stimulated photon in exactly the same state as that of an initial photon and the emission of a stimulated train in phase with an initial train are two methods of describing the same phenomenon.

Competition between the processes of absorption and stimulated emission. Thus, we have established that two different and competitive processes are possible when photons with energy $\varepsilon = E_2 - E_1$ interact with active centres (see Fig. 9): *absorption* of photons, or *stimulated emission* of additional photons. The greater the number of active centres in the appropriate initial state the higher the probability of each of these two processes is. If the number of active centres on level *1* is greater than that on level *2*, absorption becomes a more probable process. If, on the contrary, there are more active centres on level *2*, stimulated emission becomes more probable.

Consequently, the inverted population of levels *1* and *2* is a necessary condition for stimulated emission to exceed absorption. Absorption normally proceeds at a higher rate than stimulated emission (higher levels usually have lower population density), so that a light wave travelling in a medium is damped out. If, however, the population of active centres is

inverted, a light wave may be amplified in an active medium. Obviously, this wave must be coherent and have the frequency $\nu = (E_2 - E_1)/h$.

But what can produce a coherent light wave, the wave that ultimately is our goal?

The idea of selectivity for photon states. Each photon emitted by inevitable spontaneous transitions $2 \rightarrow 1$ may be absorbed or may stimulate the emission of new photons. The emission of new photons predominates in a medium with an inverted population of levels *1* and *2*. In other words, spontaneously emitted photons trigger numerous stimulated emissions.

Unfortunately, active centres emit "spontaneous" photons independently and therefore in arbitrary states. First of all, the photons differ in the direction of their momenta. It is clear that each stimulated photon is in the state of the corresponding initial photon, so that a "spread" in states of spontaneous photons results in a "spread" of stimulated photons as well. Obviously, such stimulated radiation cannot have a high degree of coherence.

Let us assume, however, that we have succeeded in creating favourable conditions for stimulated emission for some (rather few) photon states, and unfavourable conditions for all remaining states. (For the sake of brevity, we shall use the terms "favoured" and "non-favoured" states.) Then a large number of photons are generated by photons that are spontaneously emitted in "favoured" states, while the photons spontaneously generated in "non-favoured" states quite soon leave the field without stimulating appreciable numbers of photons.

Evidently, the flux of photons in "favoured" states forms highly coherent optical radiation emitted by the active medium. (In terms of Fig. 9, the radiation is generated by the $2 \rightarrow 1$ transition, levels *2* and *1* being referred to as lasing levels.)

The lower the number of "favoured" photon states, the more pronounced they are and the more "suppressed" the remaining photon states are, the better the selectivity of emission is and the higher the coherence of the emitted radiation. If it were possible to single out just one photon state, the generated light wave would be ideally coherent, that is, it would be a monochromatic plane wave with a fixed polarization.

How can we realize selectivity for photon states? Several approaches may be suggested for realizing selectivity. For instance, *directional selectivity* is achieved by preparing the active medium in the shape of a long rod with a comparatively small cross-section. Spontaneous photons with momentum parallel to the rod's axis are obviously favoured, since they can interact with a higher number of active centres and therefore stimulate an intensive avalanche of stimulated photons. On the other hand, those spontaneous photons with momenta at an angle to the rod's axis are soon leaving the active medium. The path within the active medium travelled by favoured photons may be increased still further by placing partially reflecting mirrors at the end faces of the rod. The mirrors return part of the radiation back to the active medium and therefore enhance the effect of stimulated emission for the photons that are forced to travel over longer paths in the active medium. This in fact is the idea of an

optical resonator, an essential element of a laser, which will be discussed later in the book.

The selectivity of photon *energies* is furnished by active centres with an appropriate system of energy levels. In real situations, however, the system of levels of an active centre is always richer in terms than the one shown in Fig. 9. It is possible to suppress the unnecessary levels and the transitions between them by, for instance, making the reflection coefficient of the above-mentioned mirrors frequency-dependent. Such mirrors create a selectivity of emission over photon states.

Conclusions. In order to obtain highly coherent optical radiation, it is necessary, *first*, to excite active centres and produce an inverted population of the lasing levels, and *second*, to create selectivity for photon states. The inverted population of the lasing levels of active centres is required in order that stimulated emission predominate over absorption. And selectivity is required for the effect of stimulated emission to be significant only for a few photon states (otherwise no sufficiently coherent radiation can possibly be obtained).

High coherence of laser emission is based, *first*, on the nature of stimulated emission, which makes it possible to accumulate photons in selected photon states, and *second*, on the creation of selectivity of emission in the medium, which permits the utilization of such accumulation only for a few photon states.

4. The Laser: Working Principles

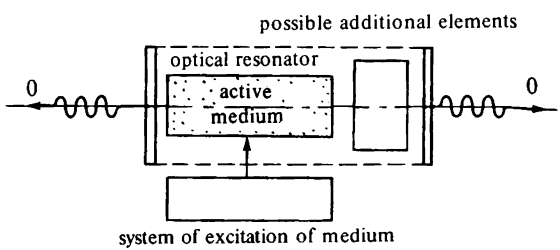
A functional schematic of a laser is shown in Fig. 10. The active medium and the additional elements are located inside the optical resonator. The resonator singles out an optical axis *OO* of the laser; the emitted radiation propagates along *OO*. Note that a laser can emit radiation both in a single direction and in two opposite directions along the optical axis.

A laser is triggered by initiating its pumping system. This system provides excitation of active centres, and an inverted population of lasing levels builds up. The optical resonator (together with some additional elements) provides selectivity over photon states. As a result, a highly coherent radiation, called the laser radiation, appears along the axis *OO*.

Active mediums and methods of excitation. The following active mediums are used in lasers:

(a) gases and mixtures of gases (gas lasers);

Fig. 10



(b) crystals and glasses doped by special ions (solid-state lasers);

(c) liquids (liquid lasers);

(d) semiconductors (semiconductor lasers).

Only *gas* and *solid-state* lasers are discussed in this book.

Normally the active medium of a gas laser is a mixture of several gases; atoms or molecules of one of them are active centres while other gaseous components serve to produce population inversion on the lasing levels of the active centres. One possible mixture, for instance, is helium and neon (neon atoms are active centres). This mixture is placed in a gas discharge tube at low pressure: neon at a pressure of about 10 Pa and helium at about 100 Pa.

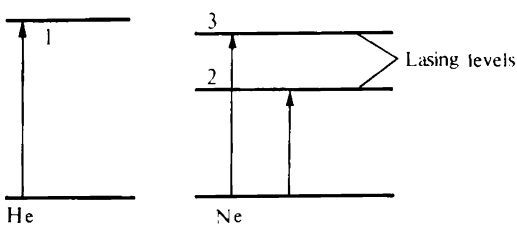
The active medium of a solid-state laser is normally a rod with a circular cross-section, doped with special ions that play the role of active centres. A classical example of a solid lasing medium is a ruby rod, 3 to 20 mm in diameter and 5 to 30 cm long. Ruby is crystalline alumina (Al_2O_3) doped with chromium ions (from 0.05 to 0.5%). Note that it is this impurity (dopant) that gives ruby its typical colour (from pink to deep red).

Systems of pumping vary, and to a large extent depend on the type of active medium. Excitation in gas lasers is realized in the simplest manner by means of electric discharge in the active medium (typically glow discharge). In this case the energy of excitation is transferred to active centres as a result of collisions with particles in the gas discharge plasma.

In the case of the helium-neon gas laser, the discharge is dc glow discharge. A simplified diagram of energy levels in neon and helium atoms is shown in

Fig. 11; the lasing levels are levels *2* and *3*. Because of collisions with electrons in the gas-discharge plasma, neon and helium atoms are excited to level *1* (helium atoms) and levels *2* and *3* (neon atoms). In the impacts, electrons transfer to the atoms a fraction of their kinetic energy (electron excitation). There is, however, another possibility to excite neon atoms to level *3*: resonance transfer of energy from excited helium atoms to neon atoms in the ground state, whereby the helium atom is de-excited while the neon atom undergoes a transition from the ground state to one of the excited levels. This energy transfer takes place because the corresponding energy levels in helium and neon (*1* and *3*) are almost at the same distance from the ground level, and also because the helium density in the mixture is sufficiently high (this prevents the unfavourable transfer of energy from neon to helium, since the probability of an atomic process is proportional, among other factors, to the number of atoms in the initial state). Taking into account collisions both with electrons and with excited helium atoms, we conclude that level *3* must be populated

Fig. 11



with higher probability than level 2; hence, levels 2 and 3 in neon atoms must be inversely populated.

Excitation in solid-state lasers is carried out by irradiating the lasing rod with light from a sufficiently powerful light source, such as a pulse lamp (optical pumping). In this case active centres are excited owing to the absorption of photons emitted by the pump lamp. Obviously, lasers with optical pumping can be considered as converters of optical radiation, converting, for instance, the incoherent radiation of a pulse lamp into highly coherent laser emission.

Optical resonator. An optical resonator is realized as a system of mirrors. In a more general sense, it includes not only a system of mirrors but also everything inside this system, including the active medium. Figure 12 schematically shows examples of optical resonators: (*a*)—a simple linear resonator; (*b*)—a coupled linear resonator; (*c*)—a ring resonator (*OO*—the optical axis of the laser, fixed in space by the system of resonator mirrors).

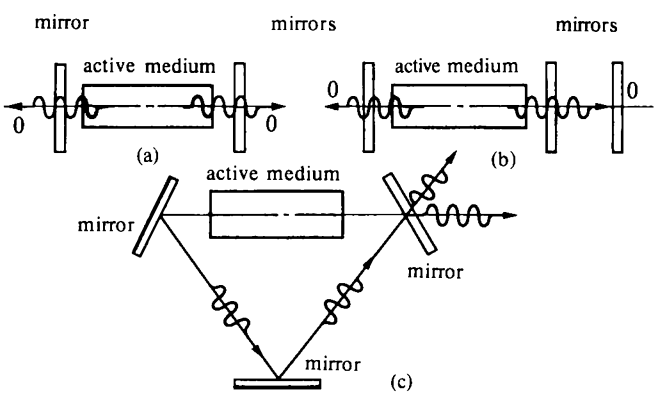
Mirrors may be coated with dielectric or metal layers. At least one of the mirrors must be partially transparent with respect to the emitted radiation; in the case of metal coating, the problem is solved by making a hole at the centre of the mirror. The technology of manufacturing laser mirrors is very complicated; a coating is deposited onto a substrate by successive thin layers. Resonator mirrors of gas lasers are usually mounted at both ends of the gas-discharge tube and are not linked to it in a rigid manner. Mirrors in solid-state lasers are typically formed on specially prepared end faces of the active-medium rod.

As mentioned above, an optical resonator defines the direction in which radiation is emitted. The processes of stimulated emission are able to compensate for absorption only for this direction.

When radiation inside a real laser is considered, it is necessary to take into account not only stimulated emission and photon absorption processes by active centres, but also many other processes causing radiation losses (such as scattering and absorption by non-active centres). Laser oscillation is sustained only if stimulated emission by active centres compensates not only for the absorption by these centres but for all other losses as well.

The concept of "favoured" photon states introduced in the preceding Section can now be elaborated. Namely, the "favoured" states are those states for which the losses are eliminated to a maximum degree.

Fig. 12



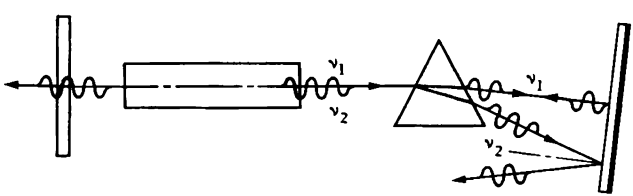
Returning to the optical resonator, we can now formulate that it singles out a direction in space for which the losses are minimized and the condition of generation is satisfied. Here we see the role of the optical resonator revealed as that of an element creating selectivity for photon states.

A prism inside a resonator. Additional elements within the optical resonator also serve to maintain the conditions of selectivity. For example, a prism inside the resonator provides *energy* selectivity.

Let us assume that the active medium generates at two frequencies, ν_1 and ν_2, simultaneously, and that the ν_2 emission must be suppressed. Various methods could be used for this purpose. One of them is based on resorting to the phenomenon of *dispersion of light*, that is, the dependence of the refractive index on the frequency of light waves.

If a prism is placed inside the resonator, the light waves with different frequencies emitted from the active medium are spatially separated by the prism (Fig. 13). If the right-hand mirror (in the figure) of the resonator is perpendicular to the propagation direction of the ν_1 wave, then the ν_2 wave is incident on the mirror

Fig. 13



obliquely and is not returned into the active medium after reflection. Consequently, only photons with energy $h\nu_1$ are returned into the active medium, while those with energy $h\nu_2$ leave the system immediately after reflection. The prism clearly realizes the selectivity effect: photon states with energy $h\nu_1$ are "favoured" while those with energy $h\nu_2$ are "non-favoured"

Note that although Figure 13 shows two light waves spatially separated by the prism, in reality there is no wave with frequency ν_2. The prism damps this wave so effectively that it is simply not emitted at all. If the prism is removed, both waves are generated; if the prism is there, the generation of one of the waves is damped out. The correct interpretation of Fig. 13 is: if a wave with frequency ν_2 appears, it is immediately suppressed by the prism.

What is the Brewster angle? When a light beam is incident on the surface of a medium at a certain angle, it is transformed into two beams, one reflected and one refracted. The first of them propagates outward at an angle γ equal to the incidence angle, and the second is refracted into the medium at an angle β satisfying the relationship

$$\frac{\sin \gamma}{\sin \beta} = n \tag{4.1}$$

where n is the refractive index of the medium.

Experiments demonstrate that reflection and refraction change not only the direction of propagation of light beams but their polarization as well. In the general case, an unpolarized light beam is transformed into partially polarized refracted and reflected beams.

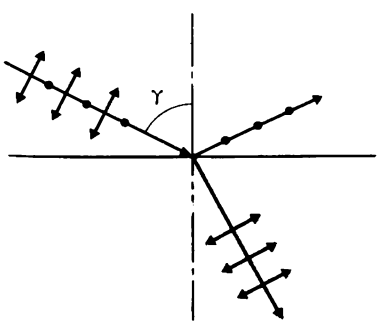
The refracted beam is polarized predominantly in the plane of incidence, while the reflected beam tends to be polarized in the plane normal to it.

If, however, the angle of incidence γ satisfies the condition

$$\tan \gamma = n \tag{4.2}$$

then the refracted and reflected beams are perfectly polarized. The angle of incidence γ satisfying the condition (4.2) is called the *Brewster angle* (Fig. 14). Dots and arrows in the figure indicate the direction of oscillation of the vector of electric field (arrows denote oscillation within the plane of the drawing, and dots — normal to this plane). It can be easily found that for an angle of incidence equal to the Brewster angle, the reflected and refracted beams are at right angles. Indeed, Eqs. (4.1) and (4.2) yield that $\sin \beta =$

Fig. 14

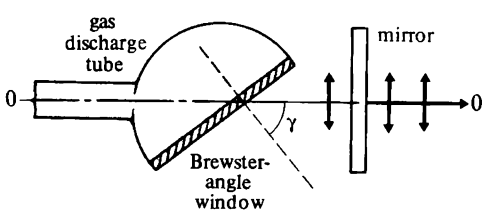


$= \cos\gamma$. Hence, angles β and γ are complementary, that is, together make a right angle.

Why do we orient the end face of the gas discharge tube at the Brewster angle? The planes of the end faces of gas-laser discharge tubes are usually tilted at an angle so that the normal to the end face and the optical axis are at the Brewster angle corresponding to the refractive index of the material of which the end-face window is made. Figure 15 gives a schematic of one end of a gas discharge tube (OO — optical axis; and γ is the Brewster angle). A similar angled window is installed at the opposite end of the tube.

The end face of the discharge tube is oriented at the Brewster angle in order to realize the conditions of polarization selectivity of emission in the gas laser. Let a non-polarized light wave be incident on the plane-parallel window (end plate) of the tube along its axis. An unpolarized wave can be expressed as a sum of two plane-polarized waves one of which is polarized, for example, in the plane drawn through the normal to the window and the tube axis (let us refer to it as the "plane of oscillation"), and the other normal to this

Fig. 15



plane. Clearly, the wave polarized normal to the "plane of oscillation" is completely reflected by the window plate while that polarized in this plane is transmitted by the plane-parallel window plate in the same direction. This means that the first of these waves is immediately shut out of the picture while the second one is repeatedly reflected by the resonator mirrors and correspondingly passes repeatedly through the active medium. This system "favours", therefore, the photon state with polarization in the "plane of oscillation" and is "unfavourable" to the states polarized normally to this plane. The emitted radiation is thus polarized in the "plane of oscillation"

Note that this device kills two birds with one stone: *first*, we obtain plane-polarized laser emission, and *second*, we eliminate losses that would be caused by reflection at the surfaces of the tube window plates. Indeed, the light wave that would have been reflected by these surfaces by virtue of its polarization is simply not emitted (Brewster windows make attenuation of this wave too high.)

Basic modes of laser oscillation. Three basic modes are distinguished: continuous oscillation, pulsed mode of free oscillation, and pulsed mode with controlled losses (Q-switching, or the so-called giant-pulse mode of oscillation). The continuous mode is typical for gas lasers, and pulsed modes are mostly employed in solid-state lasers. It must be mentioned, however, that if necessary, any type of laser can be made to operate in any of these modes of oscillation.

In the case of continuous oscillation, a laser emits a continuous light beam with constant power. This mode requires a continuous steady-state pumping of the

active medium. This pumping can be produced by steady-state discharge in a gas or by a continuously operating pumping lamp (in the case of optical pumping).

In pulsed free oscillation the emission takes the form of periodic (and sometimes irregularly repeated) light pulses each 10^{-6}-10^{-3} s long, emitted at a frequency of 10 Hz to 10 kHz. The pulsed character of emission is a result of the pulsed operation of the pumping system (a pulsed lamp, or pulsed discharge in the gas). This pumping creates population inversion on active centres periodically and for short periods only.

The pulsed mode of oscillation enables us to achieve considerable concentration of light energy, and to obtain, within short time intervals, substantial power output. If, for example, a laser emits 10 light pulses per second, each pulse 10^{-4} s long, with energy per pulse 1 J, then the maximum power achieved (referred to as the peak power) is not less than 10^4 W, while the average power of emission is only 10 W. The concentration of light energy reaches a maximum in the giant-pulse mode of oscillation.

Giant pulses. This special mode of laser oscillation will be discussed in some detail (another term: Q-switching, i.e. modulation of losses). Light pulses emitted in this mode are indeed giant, since their peak power reaches 10^9 W. For comparison we remind the reader that the biggest hydroelectric power plant in the world (Krasnoyarsk Power Plant on the Yenisei river in the USSR) has a power of $6 \cdot 10^9$ W. Such pulses, however, have very short duration, on the order of 10^{-8} s, so

that the total energy of light emitted per pulse is only about 1 to 10 J*

Giant-pulse mode of oscillation can be realized by controlling losses inside the resonator. If these losses are somehow sufficiently increased for a certain time, oscillation cannot develop. Consequently, the system of pumping builds a considerable overpopulation of the upper lasing level in active centres. If, however, the losses are then sharply brought down, the process of stimulated emission by the "favoured" photon states develops in an avalanche manner and gives rise to a short light pulse with very high peak power.

In order to control the losses in a resonator (to control its Q-factor), a so-called *optical switch* is introduced into it. In the simplest case, it is an elementary chopper that rotates and therefore periodically interrupts the light beam.

In particular, one of the mirrors of the optical resonator can be made rotating. In this case losses are small (Q is high) only during the short interval when the rotating mirror is nearly perpendicular to the optical axis of the laser; hence, this is the time when a giant pulse can be emitted.

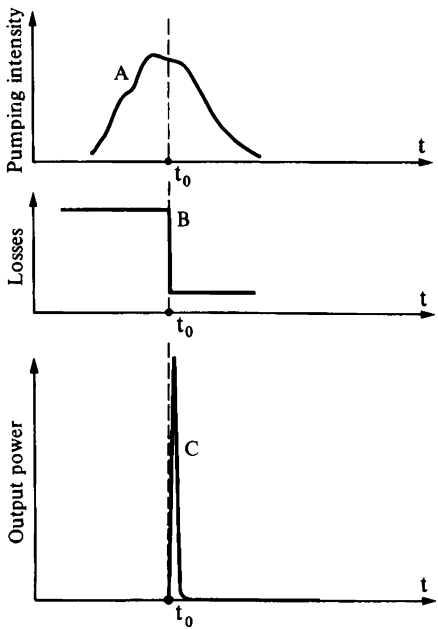
It is essential for the chopper of the beam or for a rotating mirror to be synchronized in time with pump pulses, for instance, with the pulses of a pump

*) There are also so-called axial mode-locking oscillations, which make it possible to obtain especially short light pulses 10^{-11} to 10^{-12} s long. The peak power achieved in such ultrashort pulses reaches 10^{12} W. Ultrashort light pulses are formed owing to a peculiar interference effect that redistributes light energy within each giant pulse and concentrates this energy into several ultrashort (superpowerful) pulses.

lamp: the steep decrease of losses must take place immediately after a substantial population inversion of the lasing levels is achieved.

Figure 16 shows three curves. Curve *A* plots the pumping intensity as a function of time, that is, gives the shape of the pumping pulse. Curve *B* plots a quantity characterizing losses, also as a function of time. The curve shows that losses dropped sharply at time moment t_0. Finally, curve *C* plots the dependence

Fig. 16



on time of the power of the emitted radiation; this is the shape of the giant pulse. The pulse is emitted directly after the drop in losses occurs. The figure clearly demonstrates that the moment t_0 is chosen in synchronism with the pumping pulse. The process can be repeated if losses are again increased after the pulse is emitted.

It should be emphasized that emission of a giant pulse is meaningful only if the drop in losses is very sharp. From this standpoint, rotating optical switches are not very suitable: as all other devices based on mechanical motion, these Q-switches have prohibitively high inertia. So-called electrooptical and passive (phototropic) switches, which are optical switches without moving (rotating) parts, are more effective. In electrooptical switches an external electric field modulates transparency (and consequently, losses) of the switch; in passive switches (also called photon-activated switches) losses are modulated by the generated laser emission. Passive switches are especially promising; they are discussed in detail in Chapter 3.

Finally, we note that rotating optical switches enable us to generate light pulses with power up to 10^7 W (pulse duration not less than 10^{-7} s). Shorter and, therefore, more powerful light pulses are obtained with electrooptical and photon-activated switches.

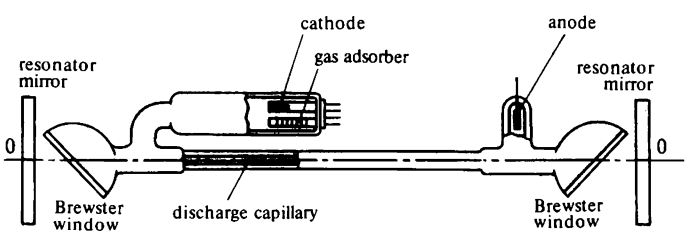
5. Lasers as Sources of Coherent Optical Radiation

This Section is chiefly for reference. A reader interested mostly in the physical side of the topic can safely skip it on the first reading.

Several typical types of lasers. The number of laser systems in use today, differing in active mediums and methods of pumping, runs into the hundreds. Only three types will be mentioned here: the helium-neon gas laser, and two solid-state lasers, the ruby crystal laser (historically the first to appear) and the yttrium aluminum garnet laser (at present the most widely used solid-state laser).

Helium-neon gas laser. Figure 17 gives a general idea of the construction of helium-neon lasers. The schematic shows the gas-discharge tube and mirrors of the resonator. The active medium consists of a mixture of helium and neon contained in the discharge capillary (internal diameter from 1 to 10 mm). The reader recalls that neon atoms are active centres, and helium is an auxiliary gas serving to produce an inverted population of the lasing levels of the active centres. The schematic clearly shows that end faces of the discharge tube are not perpendicular to the tube axis; in fact, the normal to a window plate is at the Brewster angle to the tube axis.

Fig. 17

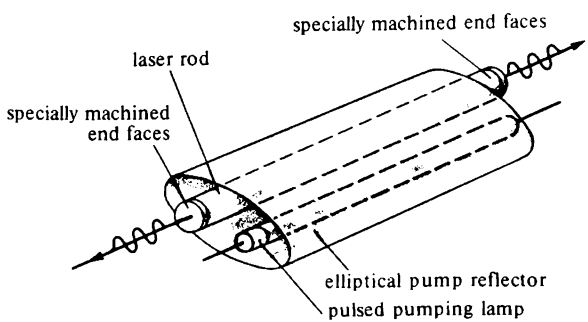


Helium-neon lasers have emission power on the order of 10 mW and an efficiency not higher than 0.1%. The lasers operate in the continuous wave mode (CW). The basic generated wavelength is 0.63 μm (red beam). The lasers may also oscillate in the infrared range at wavelengths of 1.15 μm, 3.39 μm, and some others.

A solid-state laser with optical pumping is shown in Fig. 18. The pumping lamp and the laser rod are located inside the reflector parallel to each other and pass through the focuses of the elliptical cross-section of the reflector. This achieves a high concentration of the light flux of the pump lamp on the laser rod.

Concentration of active centres in solid-state lasers is by several orders of magnitude higher than in gas lasers (10^{17}-10^{20} cm^{-3}). As a result, population of the upper lasing level is much higher and the emitted power is sufficiently high for a relatively small length

Fig. 18



of the active medium. As a rule, solid-state lasers operate in the pulsed mode of oscillation.

Historically, the first laser was the solid-state ruby laser (crystalline alumina doped with chromium ions serving as active centres). The laser emits at the wavelength of 0.69 μm; average emitting power is on the order of 1 W and efficiency is up to 1%.

A solid-state laser widely used now is the yttrium-aluminum garnet laser. The active medium (lasant) is yttrium-aluminum garnet ($Y_3Al_5O_{12}$) doped with approximately 3% neodymium ions, which serve as active centres. The laser oscillates at the wavelength of 1.06 μm (in the infrared range), with an emitting power from 10 to 200 W and efficiency up to 1%.

The wonderful laser beam. Imagine a helium-neon laser operating in a darkened laboratory. The rich red colour of the beam is a wonderful sight in the semidarkness of the room. The beam looks very unusual: no divergence (widening) is noticeable, and intensity is practically constant. One can place a number of reflecting mirrors in its way and make the beam trace a zigzag path within the laboratory. The result is magnificent: darkness crossed in all directions by bright-red filaments. If the beam diameter is magnified by means of a lens and then the beam is thrown on a screen, such as a sheet of paper, a very unusual light spot is observed: it "speckles", dark and bright spots appearing and vanishing.

The unusual behaviour of the laser beam is produced exclusively by its high degree of coherence. The first corollary is a very low divergence of the beam, and consequently, almost constant intensity as we move away from the laser. The richness of the

beam's red colouring is due to the high degree of monochromaticity of the emission.

The speckle pattern on the light spot is also caused by the coherence of emission. Light and dark specks appear because of the interference of coherent beams reflected to the observer's eyes from different points of the spot. Slight unconscious motions of the observer's head change the angle at which parts of the spot are seen and modify the conditions of interference, so that bright spots are turned into dark ones, and vice versa.

Characteristics of coherence of gas and solid-state laser emission. Different lasers emit radiation with a different degree of coherence. In order to characterize the degree of coherence quantitatively, we consider, *first*, the coherence time τ, and *second*, the angle of divergence of the light beam.

The best coherence is found in gas laser emission. Here the coherence time reaches about 10^{-3} s. This means that the wave train length (the product τc) comes up to 10^5 m (up to 100 km!). It can be mentioned for comparison that in non-laser light sources the wave train length is shorter by at least seven orders of magnitude! The angle of divergence of the gas laser emission is only one minute.

Solid-state lasers emit radiation with a lower degree of coherence. Typical coherence time is 10^{-6} s (wave train length about 100 m), and an angle of divergence on the order of 10 minutes.

The emission of solid-state lasers is less coherent in comparison with that of gas lasers because spontaneous photons in these lasers are spread in a wider range of states, and because it is more difficult to achieve high selectivity conditions for photon states.

One of the reasons for a higher spread of photons over the states lies in the difficulties involved in growing a solid active medium with sufficiently uniform optical properties and in creating sufficiently uniform optical excitation of the lasing rod. Furthermore, the "spread" in photon energies is additionally increased by a comparative abundance of energy levels in the solid state.

High selectivity of photon states is also achieved with difficulties, partly because of the impossibility of obtaining sufficiently long lasing rods. Obviously, the longer the rod (the longer the optical resonator) the better the selectivity over orientation of photon momentums is realized.

In the case of gas lasers all these difficulties are in fact absent since a gaseous medium is very uniform (homogeneous) and the gas-discharge excitation can also be made very uniform. At the same time, low density of the gas is not an obstacle to increasing the resonator's length.

The range of wavelengths "mastered" by laser technology. The available lasers emit light waves in the visible, infrared, and near ultraviolet spectral ranges. The IR range is effectively "mastered" by using active mediums operating on transitions between energy levels in molecules, while UV lasers are built on the basis of transitions between energy levels in ions.

At the present moment the range of the wavelengths is from 0.2 μm to approximately 100 μm. Highly coherent high-intensity radiation can actually be obtained on any wavelength within the indicated range. Current research effort is aimed at further broadening the "laser range", especially at lowering the short-

wavelength limit, down to the X-ray range of the spectrum. Tremendous difficulties will have to be overcome before this problem is successfully solved.

The fields in which lasers are used are multiplying. Two main approaches to the application of lasers in science and industry can be indicated: *first*, coherent optical radiation as a factor affecting materials, and *second*, coherent optical radiation as a means for the transmission and processing of data (so-called information-oriented applications).

Effects on materials. The high coherence of laser emission makes it possible to realize a tremendous spatial concentration of light power, such as 10^{13} W in a space with linear dimensions of only 1 μm. Radiation of such intensity can cut metal, produce microwelding, drill microscopic holes through diamond crystals, and so on.

Today Garin's fantastic hyperboloid has "turned into" a real laser that is widely used in industrial processes for high-precision treatment of materials.

Many complex, fine operations in surgery today are conducted not only with the traditional scalpel and lancet, but also with the laser beam. As an example, we mention the operations of welding of detached retina on the fundus oculi, penetration of blood vessels in the eye for treating glaucoma, destruction of malignant tumors, and so on.

Laser radiation may, without destroying a material, considerably change its properties and above all its optical properties (refractive index, dielectric susceptibility, transparency, etc.).

Information-oriented applications. Information can be transmitted if a coherent electromagnetic wave

(so-called carrier wave) is appropriately modulated; this means that one has to "occupy" a frequency range $\Delta\nu$ whose width depends on the character of the information to be transmitted. For instance, transmission of speech or music requires that the carrier wave be modulated by sound vibrations. The range of frequencies audible to the human ear covers the range from 16 to $2 \cdot 10^4$ Hz, so that modulation of the carrier wave must occupy a frequency band $\Delta\nu \approx \approx 10^4$ Hz (a narrower band, about 10^3 Hz, is sufficient for transmission of speech only). Transmission of images (TV videosignals) needs a much wider frequency band, about 10^7 Hz.

The required bandwidth is easy to estimate. Indeed, normal visual perception of a moving picture requires that frames be projected at a frequency of about 50 Hz. If the electron beam scanning the cathode ray tube (CRT) screen covers it in 400 lines, this means that lines are "switched" with frequency of $50 \times \times 400 = 2 \cdot 10^4$ Hz. Finally, if each line is divided into 400 "image points", the necessary frequency of electron beam modulation becomes $400 \times 2 \cdot 10^4 = 8 \cdot 10^6$ Hz.

Obviously, the frequency of modulation must be substantially lower than that of the carrier wave. The higher the latter, the wider is the frequency band that can be used for modulation. The development of coherent radiation sources in the optical frequency range makes it possible to increase the frequency band for modulation, $\Delta\nu$, approximately to 10^{12} Hz; this means that up to 10^9 telephone conversations or up to 10^5 television programmes could be transmitted simultaneously on a single laser beam. In other words, the "information capacity" of coherent radiation in the optical range is tremendous.

It should be mentioned that the high information capacity of the optical range can be fully utilized only if we solve the problem of modulation of optical radiation at frequencies on the order of 10^{12} Hz (and even higher). This is an extremely difficult problem. Modulation frequencies realizable when this book was being prepared reached "only" 10^9 to 10^{10} Hz.

Laser technology is promising not only from the standpoint of *transmission* of data, but also with respect to data *reception* and *processing*. Lasers are successfully used today to measure distances and velocities with extremely high accuracy, to detect internal stresses and defects in the framework of non-destructive structural control techniques, to detect weak underground shocks, to measure drift of continents, and so on. Lidars, that is, optical locators, are used on an ever increasing scale. The methods developed in optical holography (Sections 6-11) are being applied more and more to process information by means of coherent optical radiation (Sec. 11).

Problems in the near future. The fields to which lasers make a significant contribution are growing in parallel with the progress in the domain of laser technology. In some fields lasers already have a high standing, in other areas only first applications have appeared, while in still other ones the possibilities of their application are only at the appraisal stage.

One of the most pressing and important problems, and one that is being solved now, is that of controlled nuclear fusion. Very serious arguments point to the possibility of successful solution of this problem if superpowerful laser pulses are used. This approach is under scrutiny in the USSR, USA, and other countries.

Another important problem is connected with the need to increase the speed of operation and the memory capacity of computers. This is demanded by the needs of scientific and industrial progress, and by the necessity to manage industry and economy of the country as a whole.

It can be expected that we shall witness a qualitative breakthrough in this direction very soon; it is anticipated that purely electronic techniques of data processing will be replaced by a combination of electronic and optical methods, that is, traditional electronics will be replaced by *optoelectronics*. Such a transition is only possible after a massive penetration of lasers into cybernetics and computers. The first steps in this direction are already a reality.

The prospects promised by coherent optics are spectacular. In the chapters to follow the reader is invited to make a better acquaintance with some of the most exciting and promising branches of this novel domain of modern physics.

Optical Holography

6. Formation of Optical Images

Let us place an arbitrary object in front of a screen and illuminate it. Light beams reflected by each point of the object illuminate all points of the screen (Fig. 19); the beams scattered by different points of the object are "tangled up" As a result, the screen is illuminated more or less uniformly. An image of the object can only be obtained if the light beams are somehow "disentangled" and the patterns of rays scattered by the object are "ordered"

Pinhole camera. The light rays are easily ordered by placing an opaque sheet with a pinhole in it between the screen and the object. This is the idea of

Fig. 19

the *pinhole camera*, which consists of a dark box with a pinhole in one of the walls; the wall opposite to the pinhole serves as a screen.

Each point of the object sends only a narrow light beam through the hole, which produces the image of this point on the screen. Hence, an inverted image of the object is seen (Fig. 20).

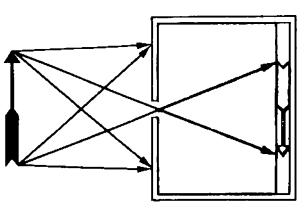
For an image to be sharply defined, the pinhole must be very small, of a diameter from 0.5 to 0.1 mm. Obviously, the smaller the pinhole the lower the irradiance of the image is. This is the reason for the dark box, and this is why the object must be very brightly illuminated.

Pinhole cameras utilize a very insignificant fraction of the light flux reflected by the object, and therefore were never widely used as image projectors.

Lens systems. Photography. The most widespread method of obtaining optical images is based on *lens systems.*

Figure 21 shows how a lens placed between an object and a screen collects, on a single point of the screen, all the light rays that are scattered by this

Fig. 20



point of the object onto the whole surface of the lens. The image is formed by a much larger light flux than in the case of the pinhole camera. One of the factors determining this light flux is the lens diameter.

The photographic method is based on lenses. The image of the photographed object is recorded on photographic plates or films. Absorption of light in the light-sensitive layer results in chemical reactions and in formation of a latent image, transformed to a visible one by the process of developing. The image is then fixed and may be kept for indefinitely long periods.

A screen with a light-sensitive layer that, after an appropriate treatment, retains the image of an object for long periods of time is referred to as a *photodetector.*

Holography. This is a fundamentally new method of recording optical images; the progress in holography during the last decade was tremendous. The origin of the term will be clarified later.

Let us analyze Fig. 22. An object is placed in front of a photodetector and illuminated with a coherent light wave emitted by a laser. The light wave reflected

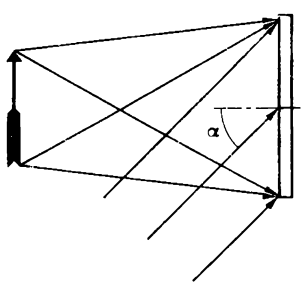
Fig. 21

by the object (the object wave) falls on the screen. In addition, another wave (also coherent, emitted by the same laser) is also directed at the screen. The auxiliary light wave shown in the figure is a plane wave incident on the screen at an angle α.

The figure shows that the light rays scattered by the object are still "tangled": each point of the object sends rays to all points of the photodetector (screen). It is essential that the experimenter makes no attempt to "disentangle" the light beams by additional devices (such as a pinhole screen or a lens) and does not try to create "order" in the pattern of scattered rays. No wonder that the photodetector with the image developed looks like an inadvertently exposed film; even the sharpest eye fails to detect the slightest trace of the object's image.

This seemingly "spoilt negative" possesses, however, one spectacular property: it has memorized the encoded (enciphered) image of the object. The decoding of this cipher, that is, the visualization of the image, is straight-

Fig. 22



forward. The image is reconstructed by illuminating the "negative" with a coherent wave identical to that used as the auxiliary one. In the case of question, the wave must be plane and be incident on the screen at an angle α.

Once such a wave is directed at the "spoilt negative", its surface miraculously becomes transparent and the observer finds himself looking at the object's image as if through a "window" (its area is that of the photodetector).

Recording and reconstruction of a hologram. This "negative" is called the *hologram* of the object. It records the interference pattern formed by the superposition of two coherent light waves: one scattered by the object (the *object*, or *signal* wave) and the other, by auxiliary (the *reference* wave). Holography is in principle an interference-based technique; it is logical therefore that light waves with a high degree of coherence are required for its realization.

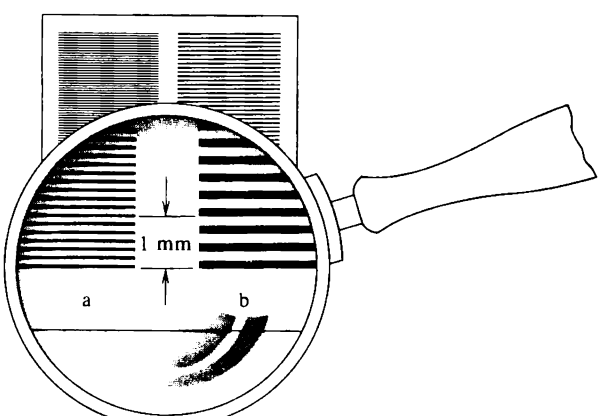
The interference pattern recorded by a hologram has a very detailed fine structure. Two interference fringes on a hologram may be spaced by only 0.001 mm, so that this fine structure is not resolvable by the naked eye. Obviously, the spatial resolution of the photodetector material must be sufficiently high for recording such detailed patterns. *Spatial resolution* is measured by the maximum number of parallel lines per unit length (usually one millimetre) that can be distinguished on a material [Fig. 23 illustrates two cases of different spatial resolution: in case (*a*) it is twice as high as in case (*b*)]. The recording of holograms requires that the spatial resolution of materials be not worse than 1000 lines per millimetre.

Information about the object is recorded on a hologram as an interference pattern. When a hologram is illuminated with a coherent light wave identical to the reference wave (the *readout* wave), this wave is diffracted by the system of interference fringes that is fixed on the hologram and acts as a sort of a diffraction grating. This diffraction reconstructs (makes observable) the object's image recorded on the hologram.

Note that the condition of identity of the readout and reference wavefronts does not include the identity of wavelengths. Recording and reconstruction of a hologram can be carried out at different wavelengths. The effect will be a change in the scale of the reconstructed image.

We see that holography, in contrast to earlier methods of image formation, is a two-stage process. In

Fig. 23



the *first* stage a holographic "photo" is taken (the hologram is *recorded*); in the *second* stage the image of the object *is reconstructed* from the hologram (the *readout* of the image). The formation (recording) of the hologram is based on the *interference of light waves*, while the reconstruction is based on the *diffraction* of these waves.

In conclusion, we want to underline an important feature of the holographic method of image formation. It consists in the fact that the "disentangling" of light rays necessary for the reconstruction of the image (*a*) is carried out only after the image has been recorded on the photodetector, and (*b*) is in a certain sense automatic, since this "disentangling" is realized by the readout light wave.

7. Holography: Elementary Examples

Example 1. The process of hologram recording is illustrated in Fig. 24*a*. Both light waves in the figure are monochromatic plane waves with identical frequencies, the reference wave propagating normal to the hologram plane and the object wave at an angle α to it.

Interference of two monochromatic plane waves propagating at an angle α was discussed in Section 1. By using equation (1.7), we conclude that a hologram records an interference pattern formed by a system of equidistant parallel fringes spaced by

$$d = \frac{\lambda}{\sin \alpha} \tag{7.1}$$

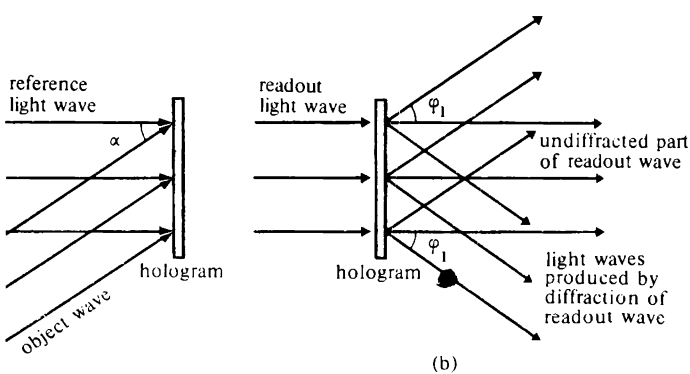
Photoprocessing of the hologram (developing, fixing, washing) yields a plate with alternating transparent and opaque parallel fringes. Such a plate can be regarded as a *diffraction grating* with period d calculated by Eq. (7.1). This period is approximately 0.001 mm; the slits on such a grating can only be seen if magnified (Fig. 25).

This hologram corresponds to an object reflecting the plane wave. A plane mirror of sufficient dimensions could serve as such an object. It can be said, therefore, that in this example we are dealing with the hologram of a plane mirror.

The process of hologram reconstruction is clarified in Fig. 24*b*. Diffraction of a wave on a diffraction grating is described by the formula

$$d \sin \varphi_k \quad k\lambda \tag{7.2}$$

Fig. 24

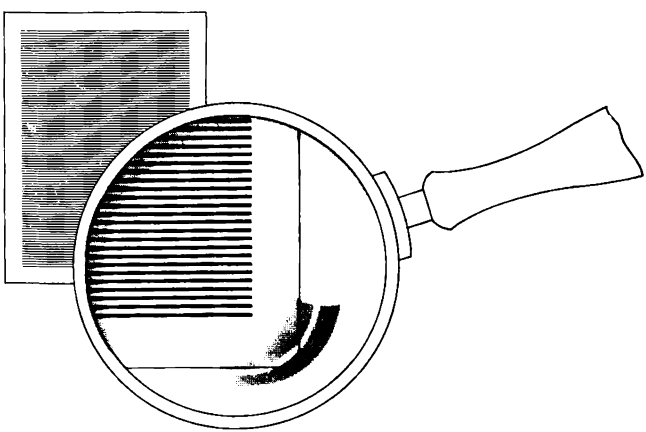


where $k = 0, 1, 2,$ d is the grating space; and φ_k denotes the angles between the normal to the grating plane and the directions at the so-called principal diffraction maximums (directions in which diffracted waves propagate). The value $k = 0$ corresponds to the undiffracted wave; $k = 1$ corresponds to two main diffracted waves. We neglect the remaining diffracted waves ($k \geqslant 2$) because of their low intensity. According to Eq. (7.2), the following is true for main diffracted maximums:

$$d \sin \varphi_1 = \lambda \tag{7.3}$$

The angle of diffraction equal to the angle of interference. Let us refer to angle φ_1 at the first principal diffraction maximum as the *angle of*

Fig. 25



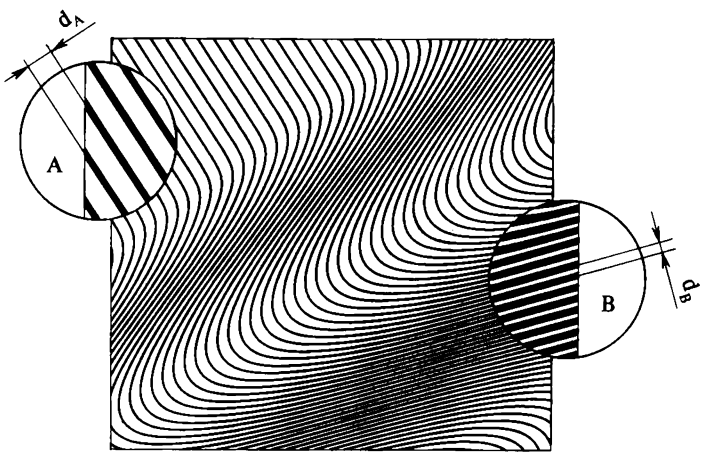
diffraction, and to angle α between the directions of two interfering waves as the *angle of interference*. By comparing Eqs. (7.1) and (7.3) we obtain that the diffraction angle is equal to that of interference:

$$\varphi_1 = \alpha \tag{7.4}$$

Obviously, in the general case the angle of interference varies over the area of the hologram. The angle of diffraction varies correspondingly. The essential point is that the equality of the two angles holds in any point of the hologram.

Note that, strictly speaking, instead of a single

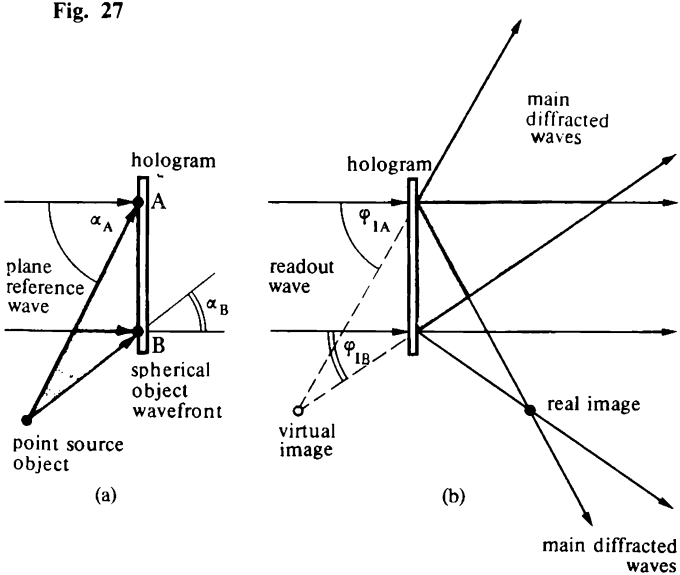
Fig. 26



point, one must consider at least a small area on the hologram. Whatever the complexity of the interference pattern, each small segment of the pattern can be regarded as a diffraction grating with a definite grating space. Figure 26 shows a fraction of an interference pattern recorded on a hologram. It illustrates that the grating space close to point A is d_A, while that in the vicinity of B is d_B.

Example 2 (point source hologram). The process of hologram recording is shown in Fig. 27*a*. The hologram reconstruction is shown in Fig. 27*b*. Figure 27*b* illustrates the rule stated above: the angles of

Fig. 27



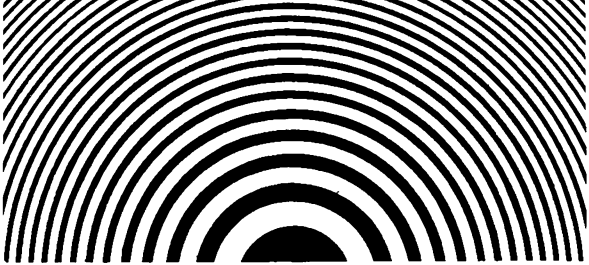
diffraction and interference are equal at any point of the hologram. Two points, A and B, are selected in the figure; hence, $\varphi_{1A} = \alpha_A$, and $\varphi_{1B} = \alpha_B$.

Figure 27 demonstrates that two images are reconstructed from a hologram: one virtual and the other real. The observer finds the virtual image in the same place with respect to the hologram as was occupied by the object in the process of hologram recording. The real image is, in this case, symmetrical to the virtual one.

The system of interference fringes in a hologram of a point object is more complicated than in the first example: the fringes are not straight and the spacings between them vary as a function of the angle of incidence of the object light beam. Figure 28 illustrates what the pattern is like.

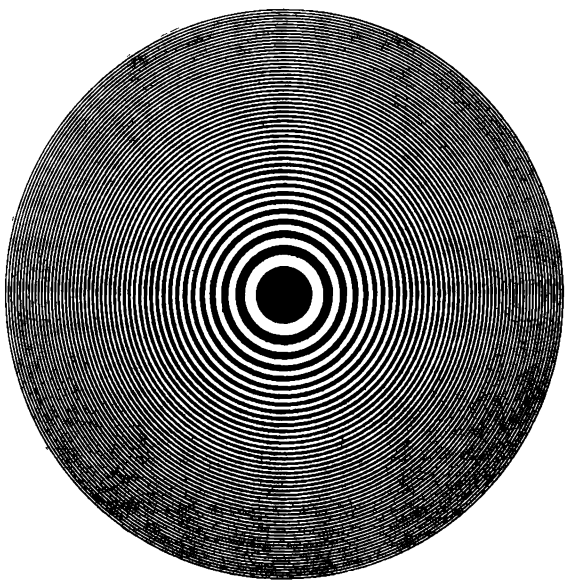
A hologram of a point object is the Fresnel zone plate. So-called *Fresnel zone plates* are a very familiar

Fig. 28



object in optics. In the simplest case a Fresnel plate is a system of alternating transparent and opaque rings whose width diminishes, according to a special formula, as their radius increases (Fig. 29). The Fresnel zone plate is an interference device that in principle is similar to the diffraction grating. Let a monochromatic plane light wave be incident on such a plate. Interference of different waves diffracted by the annular slits of the plate focuses the light at a specific point behind the plate. Hence, a zone plate is in fact

Fig. 29



a planar (two-dimensional) equivalent of a converging lens. This property of Fresnel plates is very well known, but wide use of these plates in optics was impossible because of difficulties encountered in manufacturing them (the number of rings necessary to have a good "lens" is quite high).

Holography pointed to a very attractive method of sufficiently simple production of high-quality Fresnel zone plates. It was found that the Fresnel zone plate is a *hologram of a point source (point object).*

The hologram illustrated in Fig. 28 is not really identical to the Fresnel zone plate, which can be obtained if the conditions of recording are modified: the point object must be placed as shown in Fig. 30. In this case the hologram will take the form shown in Fig. 29.

Let us denote by L the distance from the object to the hologram, and by r the distance from the observation point to the centre of the hologram. If $L \gg r$, then Eq. (7.1) yields a formula for the width of interference fringes:

$$d = \frac{\lambda}{\sin \alpha} \approx \frac{\lambda}{\tan \alpha} = \frac{\lambda L}{r} \tag{7.5}$$

Fig. 30

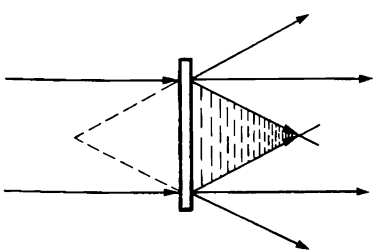
The width of the rings is thus inversely proportional to their radius.

The Fresnel zone plate as a planar equivalent of a lens. We have mentioned above that the Fresnel zone plate is a planar equivalent of a lens. This is easily proved if the zone plate is treated as a hologram of a point object.

Let us illuminate the hologram of Fig. 29 by a readout wave. Two diffracted light waves are produced, one giving a virtual image of the point object and the other forming a real image (Fig. 31). The wave forming the real image is shown in the figure by dash hatching. It is apparent that the diagram in Fig. 31 corresponds to a situation in which a plane light wave, incident on a converging lens, is focussed into a point. A hologram of the type of Fresnel zone plate possesses, therefore, the focussing properties of a lens.

The above example allows us to draw an important conclusion: in addition to serving as a "store" of the encoded image of the object, a holo-

Fig. 31



gram may be used as a "transformer" of light waves. A hologram reconstructs not just an image of an object, but a light wave with an appropriate wavefront. In other words, holograms may be used as planar equivalents of various optical elements (lenses among them) and their combinations.

Since a hologram is merely a pattern on a plane (even if a very complicated one in the general case), it can in principle be artificially reproduced. Not only the Fresnel zone-plate pattern but much more complicated interference patterns could be produced, giving us planar equivalents of the most varied optical elements.

A hologram reconstructs the object wavefront and not the object's image. Let us go back to the examples shown in Figs. 24 and 27. Among the waves produced behind the hologram in both cases of image reconstruction there is always a light wave identical to the object wave.

We have mentioned above that the hologram readout phase reconstructs not only the image but the object wave as well. This fact deserves special attention.

Take a light wave propagating away from the object (Fig. 27*a*). An observer sees the object if this wave is recorded by his eye. Let us record the object on a hologram, and then remove the object. Illumination of this hologram by a readout light wave generates several waves, among them the same object wave. And, although the object was removed, the light wave scattered by this object is reconstructed. Consequently, an observer receives the image of the object as that illuminating the object itself.

It would be more correct, therefore, to say that a hologram stores not just an encoded image of the

object but the object light wave. It is said sometimes that recording of a hologram "freezes" the object wave, while the reconstruction "unfreezes" it.

Transition from a point object to a three-dimensional one. It is evidently difficult to discuss the difference between a point object and its point image. Three-dimensional (actual) objects have to be considered.

It is fairly obvious that the fact itself of hologram reconstruction (reconstruction of the object light wave) does not depend upon whether an object is three-dimensional or not. Therefore we conclude that if it is three-dimensional, the observer looking at the reconstructed image does not see a two-dimensional picture (as in photography) but something three-dimensional and realistic; in other words, he sees something looking very much like the object did in the process of the hologram recording. If the observer tilts his head, he notices other objects behind the first one, or new details that were not noticeable before. This means that the observer receives a quite realistic three-dimensional picture.

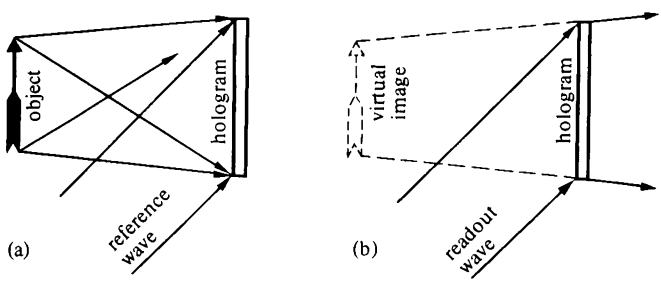
It is normal to say that a hologram reconstructs a three-dimensional image of the object. This is undoubtedly true. However, this statement is somewhat incomplete and "underestimates" the possibilities of holography. A hologram reconstructs the real object wave, which is more than an image, even a three-dimensional image. A real light wave is something one can operate with (for instance, it can be made to interfere with another wave) while an image can only be recorded.

Two arrangements of hologram reconstruction. The transition from a point object to a 3-D object can be

realized, of course, by establishing a correspondence between the 3-D object and a set of, for example, three points of the image. This evidently demonstrates that an object reconstructed by a hologram is three-dimensional. We find, in addition, that only the virtual image has all the features of the real object; the real image looks as if it is "turned inside out", that is, points farther from the observer are imaged closer to him. Such images are called pseudoscopic.

If the reconstruction geometry is changed so that the readout wave is directed opposite to the reference wave, the real image is an exact copy of the object while the virtual image becomes pseudoscopic.

Fig. 32



Two geometric arrangements are therefore possible for reconstruction of a hologram: (1) the readout wave and reference wave are identical (the arrangement was used in all the examples discussed so far); (2) the readout wave is inverted with respect to the reference wave. In the first arrangement one operates with the virtual image, and in the second — with the real one. Figure 32 shows: *a* — geometry of hologram recording; *b* — normal (ordinary) reconstruction geometry; *c* — inverted reconstruction geometry.

8. Holographic Laboratory

Typical arrangement for hologram recording. Different geometric arrangements can be used for recording a hologram. One of these arrangements (with bilateral illumination of the object) is shown in Fig. 33. Laser is a source of all the light waves used in this arrangement; semitransparent mirrors split the laser beam twice thus forming two waves illuminating the object and the reference wave.

In essence, holography is a lensless method of image formation, although lenses are used in the arrangement shown. They play an auxiliary role (serve as light beam expanders).

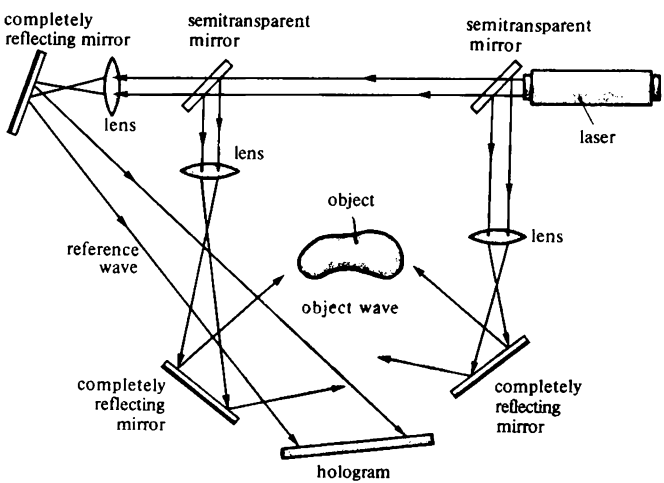
Laser. The laser is an irreplaceable instrument in a holographic laboratory. Indeed, sufficiently high coherence is the basic requirement for the light source. The required degree of coherence of the laser radiation is determined by the type of the object and the choice of the geometric arrangement of hologram recording.

The following condition must hold:

$$L \ll \tau c \tag{8.1}$$

where τ is the coherence time, and L is the maximum path length difference for two light beams propagating in the chosen arrangement from the laser to the hologram. In order to estimate L, one must take into account all possible beam trajectories in the chosen arrangement and to trace them, including all reflections and refractions on the path of the beam from the laser to the hologram. Condition (8.1) has the same meaning as condition (3.1): interference of two light beams requires in both cases that the path difference for the

Fig. 33



two beams reaching the screen be smaller than the wave train length.

Another requirement for the laser emission is that its intensity be sufficiently high. Together with the sensitivity of the light-sensitive material used to record the hologram, this characteristic determines the required time of exposure.

Holography often uses gas lasers (as a rule helium-neon lasers) operating in the continuous wave (CW) mode. Although their coherence is high, the emitting power is comparatively low (see Sec. 5). Consequently, the time of exposure has to be large, so that moving objects cannot be holographed.

Holography also employs pulsed solid-state lasers. Their emission is less coherent than that of gas lasers, but they have high peak power, which makes it possible to cut down the exposure to the light pulse length (for instance, to 10^{-3} s). Pulsed illumination is used to holograph moving objects and to record the development of a process in time.

It should be mentioned in conclusion that the requirements for the coherence of laser emission can be relaxed by reducing the depth of field recorded on the hologram and by better compensation of the path length differences of the light beams involved.

Measurement bench. The laser, object, hologram, and all necessary optical elements used in the system (reflecting mirrors, splitter plates, lenses, prisms, etc.) are fixed in predetermined positions on a measurement bench, which normally consists of a massive steel plate of sufficient area (for example, 2×2 m).

If a hologram is recorded by a gas laser, the measurement bench must satisfy very severe

requirements with respect to possible displacements of the elements during the exposure: the positions of the elements must be fixed to within 1/4 of a wavelength (in other words, with the accuracy of 0.1 μm). Violation of this condition results in blurring of the interference pattern on the hologram.

It might seem that once the elements are fixed, nothing can cause their displacement. This is wrong. The elements shift because of vibrations of the walls, floor, and building foundation. Any city is rich in causes for these vibrations: traffic, large industry, and so on. As a rule, such vibrations are not felt; nevertheless, they are a reality, and in spite of being seemingly negligible, are quite capable of breaking the severe requirement of immovability of elements on the holographic bench.

In order to eliminate vibrations, optical elements are mounted on sufficiently heavy foundations with shock absorbers (for instance, the steel plate may be placed on air-filled inner tubes of car wheels).

Holographic recording materials. It was mentioned above that materials for hologram recording must resolve not less than 1000 lines per millimetre, which means a rather high spatial resolution.

The reader will recall that photoemulsion consists of minute grains of silver bromide suspended in a transparent gelatin layer. Consequently, a developed image consists of separate "spots" (is built of specific "bricks"). Features of the image smaller than these "spots" are therefore indistinguishable. It is thus clear that increasing the resolution calls for photographic materials with finer grain structure. It should always be kept in mind that diminishing the grains invariably

entails impaired sensitivity (indeed, a photon absorbed by emulsion affects a grain as a whole, so that the larger the grains are the smaller the number of photons required to form an image). The development of photographic materials for holography with high resolution and at the same time high sensitivity is not an easy technical problem.

Photographic films now in use in holography have a resolution of 1500 to 2000 mm^{-1} and sensitivity on the order of 10^{-2} J/cm^2 There is also an organic photographic material, so-called photoresist, enabling one to achieve a resolution of 3000 mm^{-1} at a sensitivity of 10^{-2} J/cm^2 Experimental photographic plates were reported with especially fine-grain structure, realizing a resolution of 5000 mm^{-1}

Can a hologram be erased? One shortcoming common to all photographic materials is that they are not meant for reuse. A hologram recorded on a photographic plate cannot be erased and a new hologram cannot be recorded in its place. As a result, photographic materials are classified as *irreversible* recording mediums.

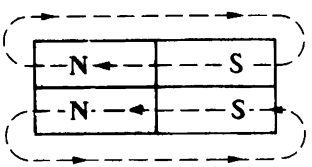
But are there any *reversible* recording mediums that allow erasing and repeated recording of holograms? Yes, such mediums exist. Let us consider some examples.

Holograms on magnetic tape. Let a laser beam fall on some point of a magnetized ferromagnetic film and heat the illuminated spot to a temperature above the Curie point. The film in this spot undergoes a transition from a ferromagnetic to paramagnetic state and loses its magnetization. Once the laser illumination is removed, the spot in question cools down, returns to

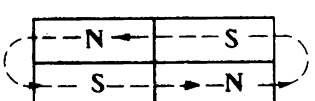
ferromagnetism, and is immediately magnetized by neighbouring regions that were not irradiated. It is essential that the direction of magnetization of the illuminated area is now opposite to that of the surrounding regions. The reason is clear from the following model situation: two rectangular bar magnets placed alongside on a slippery horizontal surface. This system is unstable while identical poles of the magnets are adjacent (Fig. 34*a*), and stable only if opposite poles are in contact (Fig. 34*b*), because in the second configuration the magnetic field lines are closed.

We thus obtain that illumination of a ferromagnetic film results in flipping the magnetization direction in places with sufficient irradiance. This process may be used to record a pattern, for instance, a pattern formed by interference fringes.

Fig. 34



(a)



(b)

In order to erase the image recorded on a film, it is sufficient to switch on an external magnetic field and magnetize the whole film in a chosen direction, after which the film is again ready to memorize a new hologram.

The time required to obtain a hologram on a photographic material is determined by the time of the photographic treatment (developing and fixation of the image). In the case of a magnetic film, however, this is the period of time required to heat the film and switch its direction of magnetization. This time can be very short, less than 10^{-7} s.

Obviously, ferromagnetic films most suitable for hologram recording are those with a relatively low Curie point. For example, a manganese-bismuth film with the Curie point of 180 °C is used (spatial resolution 1000 mm^{-1}, sensitivity to light 10^{-2} J/cm^2).

Holograms on thermoplastic films. Thermoplastic is a name of specific transparent dielectrics that soften at comparatively low temperatures (for instance, at 50 °C). A hologram can be recorded on the surface of a thermoplastic as a relief pattern. Such a hologram remotely resembles a phonograph record.

In order to use thermoplastics for hologram recording, a thermoplastic film is applied to a substrate consisting of two layers, one of a semiconductor and the other of a transparent conductor. The arrangement of the layers is shown in Fig. 35. A hologram is recorded as follows.

First of all the surface of the thermoplastic film is uniformly charged in darkness by means of glow discharge; this creates a sort of a capacitor whose "plates" are the charged surface of the thermoplastic and the conducting layer (Fig. 36*a*).

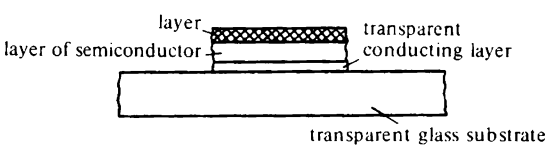
The next stage is the illumination of the whole sandwich with superposed reference and object waves. By virtue of photoconductivity, the semiconductor's resistivity increases sharply where interference maxima are formed; consequently, the distance between capacitor "plates" in these places "contracts" (Fig. 36*b*). Note that the field strength inside the capacitor remains unchanged (it depends only on the density of the surface charge). At the same time, a decrease of the interplate spacing at constant field strength results in a diminished potential difference. Hence, the potential of the illuminated areas of the surface decreases.

The surface of the thermoplastic is then recharged so that its potential is restored to the initial value over the whole surface. This concentrates additional charge on the illuminated areas (Fig. 36*c*).

Finally, the system is heated; the thermoplastic softens, and coulomb forces between charges shape the relief on the surface of the thermoplastic in accordance with the charge distribution. Cooling of the film freezes this relief (Fig. 36*d*).

Obviously, hologram formation on thermoplastics requires more time than in the case of magnetic films. The characteristic time of recording on a magnetic film is about 10^{-7} s, and reaches only 10^{-1} s in the case

Fig. 35

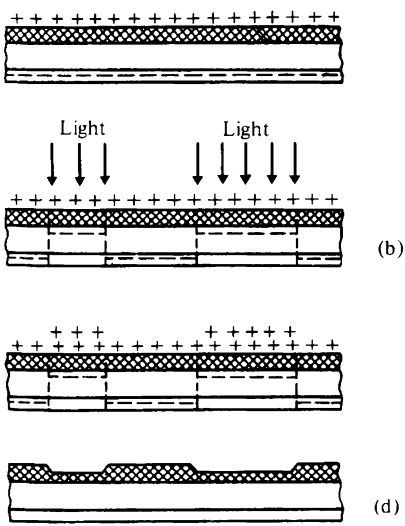


of thermoplastics. Spatial resolution in the latter can be as high as 1000 mm^{-1}

Heating is sufficient to erase the recorded relief on a thermoplastic, after which a new hologram can be recorded.

Holograms on photochromic materials. Photochromic materials are glasses doped with special impurities or organic polymers, capable of changing colouration or transparency as a result of irradiation with light in a specific frequency range (usually in the UV or short-wavelength range of the visual part of the spectrum). This is related to the transfer of electrons between

Fig. 36



impurity atoms which results from absorption of radiation.

In order to restore the initial state in the material, it has to be either heated or irradiated with light of lower frequency. Normally, photochromic materials are darkened by visual light and "bleached" by infrared light.

Photochromic materials are attractive because of their high spatial resolution, up to 3000 mm^{-1}

9. Advantages and Possibilities of Holography

A photograph and a hologram. At first glance, a photograph is preferable to a hologram. Indeed, a photograph shows everything "clearly", while nothing can be seen on a hologram. The latter has to be illuminated with laser light in order to yield an image. This, of course, means certain amount of trouble. It seems probable that those uncoded photographic images will remain preferable to encoded holographic images for ordinary purposes and in daily life. The situation is quite different in science and technology, where inconveniences caused by the encoded character of images on holograms are more than outweighed by the possibilities opened to scientists and engineers by the new technique. To be precise, it is this "encoded" character of the images that underlies the rich potentialities of holography.

A hologram enables us to reconstruct a real light wave and therefore makes it possible to manipulate optical fields (something absolutely impossible in photography). In addition, an image reconstructed by a holo-

gram differs from a photograph in its three-dimensionality, realistic and true-to-life nature.

"*Holography*" *means* "*complete recording*" A light wave can be considered as a carrier of information that is "recorded" in terms of wave parameters. It is convenient to distinguish between the information "recorded" (contained) in the wave amplitude, and the information contained in its phase (its wavefront shape). Correspondingly, *amplitude* and *phase* information are distinguished.

Let us discuss the retrieval of information from a light wave, and the recording of the information carried by this wave in a selected medium.

Let us emphasize first of all that any photodetector records the intensity of an incident light wave, that is, its response is proportional to the squared amplitude of the wave. This means that a photodetector retrieves only amplitude information and phase information is thereby lost.

If we want to retrieve and record by a photodetector not only amplitude but phase information as well, we have to show "cunning", namely we must make the light wave carrying information interfere with an auxiliary light wave. Our "ingeniousness" lies in the fact that the amplitude of the resultant light wave depends also on the relative phase of the original and auxiliary waves. Therefore, the photodetector recording the intensity of the resultant wave at the same time records not only the amplitude but also the phase information of the analyzed wave.

The method of holography is based on just this "stratagem" By using the interference of waves, we can retrieve from the object wave and record on a photo-

detector practically all the information about the object, including both amplitude and phase components. It is not accidental that the term "holography", translated from its Greek roots, means "complete recording"

Let us emphasize that if a hologram did not memorize nearly all the information carried by the original object wave, the reconstruction of the real object wave from the hologram would be impossible in principle. The process of reconstructing the object wave from a hologram indeed resembles a sort of "unfreezing" of the light wave.

Reliability of holographic storage of information. A group of people wishing to keep a momento of a meeting had first a photograph and then a hologram of the group taken. The two records of the event were stacked together and put away for long-term storage. A small fire broke out one day. Both the photograph and the hologram were saved, but both lost about one fourth of their area. Some time later one of the members of the group had to be identified by documents. When the group photograph was extracted, the face in question was missing; presumably, it had been on the burnt portion. The hologram was resorted to. It was illuminated by a laser beam, and gave the image of the whole group, including the person to be identified. The hologram proved more reliable than the photograph.

This scenario was invented in order to illustrate the following: destruction of a portion of a photographic image results in an irrepairable loss of information corresponding to a part (or parts) of the object; on the contrary, destruction of a part of the corresponding

hologram affects the reconstructed image in an absolutely different manner. Indeed, Fig. 27*a* demonstrates that the information about a point object is recorded over the whole area of the hologram. Obviously, this is true of any point of a real object: each such point is recorded on the whole hologram and not in one of its points (the situation realized in photography).

Consequently, destruction of a part of a hologram does not erase a specific portion of the image in the reconstruction. Practice shows that up to 9/10 of the area of conventional hologram can be removed without appreciable loss; the only result is a diminished resolution (sharpness) over the whole image. This means loss of fine details in the image as a whole.

Actually, this is not surprising if one recalls that a hologram reconstructs not just the object's image but the object wave. The area of the illuminated hologram determines the area of the reconstructed wavefront. Contraction of the illuminated area is equivalent to contraction of the wavefront (as if we were regarding a remote scene through a gradually contracting window). A comparatively small loss in the wavefront area does not affect the quality of the reconstructed image; greater reduction of the area worsens the resolution.

A hologram is thus a very reliable method of data storage. It is not impossible that in the future the most valuable information will be stored in holographic form.

Information is recorded on a hologram as an interference pattern, that is, in an encoded form, and it can be decoded only by a coherent light wave with exactly the same wavefront as in the reference wave. Consequently, the shape of the wavefront represents the

key without which the code cannot be broken and the hologram read out. Even the most ingenious decipherment specialist will fail if the wavefront shape is unknown (and this shape may happen to be very unusual).

Information capacity of the hologram. Is it possible to record several photographic images on the same plate? In principle, yes; but who wants a photo with several superposed images? Superposition of several printed pages would be just as pointless. This reflects the restrictions inherent in the photographic method of data storage.

Holography eliminates these restrictions thus forcing us to correct our habitual concepts. The same hologram may contain a number of consecutively recorded scenes (interference patterns) that can be reconstructed independently. Indeed, reconstruction of an image recorded on the hologram requires employing a readout wave with a wavefront structure identical to that of the reference wave and incident on the hologram in exactly the same manner.

Assume, for the sake of simplicity, that the reference wave is plane. Let us record on the hologram different scenes each time changing the angle at which the reference wave is incident on the holographic plate (this can be done, for example, by rotating the hologram in the reference beam). Evidently, reconstruction of a specific scene only requires that the hologram be properly oriented with respect to a plane readout wave. The admissible number of recorded images on the same hologram rises considerably if we take into account the possibility of changing not only the hologram orientation with respect to the reference wave but also

the possibility of changing the wavefront shape of this wave.

Estimates indicate that a single hologram with an area of about 100 cm^2 may contain (under the condition of unhindered reconstruction) at least one volume of the Greater Soviet Encyclopaedia (or Encyclopaedia Britannica, for that matter)! This points to the extremely high information capacity of holograms.

Taken together, this high information capacity of holograms and high reliability of hologram storage enable us to forecast that in the future book depositories may be replaced by hologram depositories. Instead of bulky volumes, which in addition are easily harmed, we might use miniature cassettes with holograms.

A moment stopped dead. A photo shows a diver suspended in the air several metres above the water in an unbelievable pose. Another photograph gives a magnified view of space filled with moving dust particles; the specks are frozen in the positions in which they were caught at the moment of exposure. It is said that photography is capable of "stopping time"

Let us assume, however, that we want to find out the position of the diver's left hand. Alas, it is not visible in the photo. We want to scrutinize in greater detail the dust specks in the background, but cannot because they are blurred. Obviously, there is nothing to be done about it.

Now suppose that instead of having been photographed, the diver and the scene with moving dust particles were recorded on holograms. The reconstruction in continuous laser light produces a real

light wave identical to that scattered by the object at the recording stage. And the visual perception shows the diver who seems actually hanging in the air. The same is true of the dust-particle scene. Now we can look at the diver from different viewpoints, can bring either the nearest or the farthest dust specks into the focus of a microscope, look at them from different observation points, and so on.

It is thus clear that in contrast to a photograph, a hologram, "stopping time", enables us to extract much more complete information about the object at a given moment of time. A hologram records and continuously reconstructs the structure of optical fields that existed only during the exposure. Consequently, holography provides a unique opportunity of multi-purpose processing of optical information when the actual optical fields in question are already replaced by subsequent fields (post-experiment data processing).

Holography and data processing. Optical holography is in fact a little more than fifteen years old. Nevertheless, specialists in the field of data processing devote much of their attention to holography. Let us consider some of the possible applications of holography in this field.

Pattern recognition. Pattern recognition is an important problem in cybernetics. How can we recognize a specified letter in a text? How can we select a flawed element in a set of seemingly identical parts? How can we recognize an anticipated signal in an ensemble of different signals at an instrument's input? These are specific examples from the realm of pattern recognition.

Holography is one of the promising methods of practical solution of this problem. As an example, let us consider how to use holography in order to recognize a letter in the text.

Let us choose "T" as this letter. In order to solve the problem, we need to prepare a hologram of the size of one letter of the text; the recording must be done with the wave scattered by "T" serving as reference wave, and a wave from a bright source as the object wave. A special device shifts the hologram along the lines of the text. Each time it faces a "T", a bright flash is observed, since the wave scattered by this "T" reconstructs the image of the bright source.

Associative retrieval of information. Associative search is one of the principles inherent to human memory: we start by remembering a "detail", that is, some characteristic "feature" (associative identifier), and this makes the whole picture "surface" in the memory. In other words, an associative search is the reconstruction of a whole from an individual feature.

Holography proves to be very suitable for the technical realization of the associative data search.

Assume now that a hologram was recorded without a reference wave, and only the object wave participated. Is it possible to reconstruct the image of the object?

Do not hurry with a negative answer. Indeed, we can assume (and this is perfectly true) that the interference pattern recorded in this case is formed as a result of superposition of light waves scattered by different parts of the object. Let us consider the wave scattered by one element of the object (we will refer to it as the "characteristic feature") as the reference wave, and the ensemble of waves scattered by the remaining

parts of the object as the object wave. For instance, wave in Fig. 37, reflected by feature of the object, can be taken for the reference wave. Clearly, reconstruction of the object's image from the hologram is possible if it is illuminated by the wave scattered by the "characteristic feature" This means that it is sufficient to "present" only a fraction of the object (its identifier) in order to reconstruct the whole image. In other words, the whole is reconstructed from its element, or an individual identifier. This search is clearly associative: if the hologram contains a number of images, only the one that contains the associative identifier is reconstructed.

We ascertain, therefore, that, *first*, it is possible to work with holograms recorded without a special reference wave, and *second*, that holography beautifully suits associative data search, the development of the *associative memory* (associative storage).

Encoding and decoding of information. Let us recall that a hologram carries information in an encoded

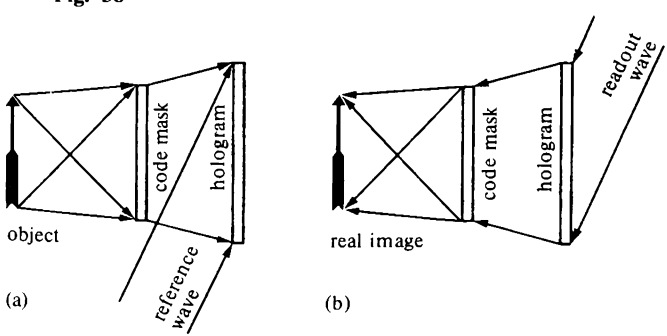
Fig. 37

(ciphered) form. Hologram recording is the encoding, and reconstruction is the decoding of this information. It is then logical to apply holography specially to encoding and decoding of data.

In order to encipher the information carried by a coherent light wave, it is sufficient to send this wave through a special plate changing the amplitude or front shape of the wave. Such plates are termed *code masks.*

Two encoding techniques are known. In the *first* method the code mask is placed in the path of the reference wave, and in the *second* method—in the path of the object wave. Decoding in the first method is realized with the arrangement shown in Fig. 32*b* (the same mask used for encoding is placed in the readout beam, and in the same position). Decoding by the second method is shown in Fig. 32*c*. We recall that this arrangement reconstructs an undistorted real image. The essence of the second method is clarified in Fig. 38. Coding is shown in Fig. 38*a*, and decoding in Fig. 38*b*.

Fig. 38



A general notion of volume holography. Until now we assumed the photodetector to be two-dimensional. As long as the thickness of the light-sensitive layer is comparable with the spacings between interference fringes, this is true. If, however, the layer thickness is much greater, the photodetector reveals specific features of three-dimensional mediums. As a result, holography is classified into *ordinary,* or *two-dimensional* holography, and *three-dimensional,* or *volume holography.* The idea of volume holography was suggested by the Soviet scientist Denisyuk.

The pattern fixed by a light-sensitive film in the region of superposition of the object and reference beams is one of interference fringes, while a light-sensitive volume fixes a system of interference surfaces. *In the first case* the result is a plane hologram, and *in the second* a volume hologram.

A system of interference surfaces of a volume hologram can reflect light waves, and this is indeed observed at the hologram reconstruction stage. The waves reflected by these surfaces interfere constructively only if their phases are in step. This means that a volume hologram manifests *selectivity* with respect to the readout wavelength: obviously, the synchronization of phases is realized only for the wavelength used to record the hologram. It becomes possible, therefore, to carry out hologram reconstruction in white light (sunlinght or light of an ordinary incandescent lamp), with the hologram "selecting" out of the continuous spectrum the very wavelength that can reconstruct the recorded image.

If several fixed wavelengths (emitted by several lasers) are used to record the hologram, reconstruction of this hologram in white light singles out the same

wavelengths. This is a method of obtaining colour images.

Holography around us. It has been mentioned above that it is hardly likely that holography will replace photographic methods in everyday situations; its maximum effect is expected in the scientific and technical domains. This does not mean, nevertheless, that people whose occupations are far from either science or technology will not feel the impact of holographic "miracles"

Even today holographic images with nearly a 360° view are used for impressive advertizing. It can be predicted that such images will be widely applied in the theatre and circus. It was experimentally demonstrated that holographic cinema is a feasible proposition. Time may come in the near future when a spectator in a movie theatre watches "live" three-dimensional holographic images: absolutely "realistic" people (as in today's theatre) amidst absolutely "real" scenery (more real than that in the theatre). It is not unlikely that our generation will see the appearance of holographic television.

In other words, holography may be expected to come into our daily life, our chores, and our leisure, and in the not very distant future at that.

10. Holographic Interferometry

The principle of holographic interferometry. As we learned earlier, holography is capable of "stopping time" at a chosen moment: the light wave that was reflected by the object at the moment of recording can be fixed and then reconstructed. We can go even fur-

ther than that: not just "stop", but "juxtapose" two distinct moments of time. This requires recording the object at two time moments on the same hologram. When this hologram is reconstructed, two light waves are produced simultaneously: one scattered by the object at moment one, and one scattered at moment two. Both reconstructed waves are absolutely real, and therefore can interfere. Observation of the interference of these two light waves is the essence of *holographic interferometry*.

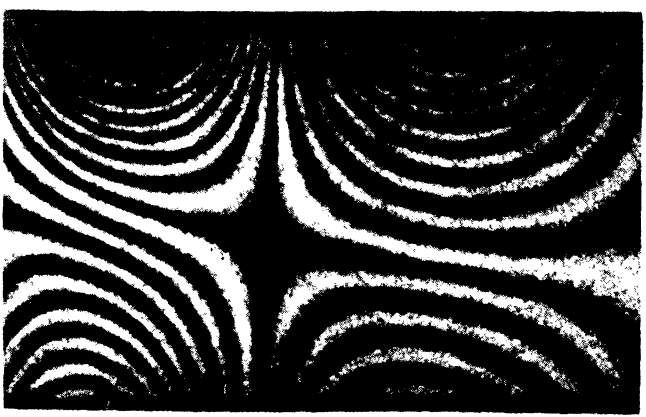
Note that holographic interferometry actually deals with several interference patterns. When an image is recorded twice and then reconstructed, three interference patterns are produced. The *first* pattern is formed by the hologram recording at moment 1; the *second* pattern comes from the repeated recording on the same hologram (the two patterns superpose on the film and are then fixed); finally, the *third* pattern is a result of interference of waves reconstructed from the first and second interference patterns. In contrast to the first two, this last pattern is usually clearly visible to the eye. It is observed as a system of interference fringes over the object's image reconstructed from the hologram. This reconstructed image is called the *holographic interferogram*. An example is schematized in Fig. 39 (a system of interference fringes covers an image of a metal plate).

The following procedure should be followed to produce such holograms. The object (plate in this case) is fixed on the measurement bench and recorded on a hologram. Note that the moment for this recording may be arbitrary since the plate is immovable anyway. Then the plate is strained by applying some mechanical stress (interesting results are obtained if the plate is

simply pressed to the bench surface in some of its points). After this the strained plate is recorded on the same hologram. Reconstruction produces two object waves, one representing the unstrained plate and another the loaded plate. By virtue of interference of these waves the resultant image contains interference fringes (something like the pattern in Fig. 39). Obviously, the extent and character of deformation of the plate over its surface can be deduced from the width and distribution of the interference fringes.

Different types of holographic interferometry. A process fundamental for the method of holographic interferometry is the interference of two light waves of which at least one is reconstructed from a hologram (sometimes the two waves are said to be compared).

Fig, 39



Two techniques are possible here. In the *first* one, both waves are reconstructed from a hologram (this version was discussed above). In the *second* version, only one of the waves is reconstructed from a hologram while the other is scattered directly by the object. Let us analyze these techniques in more detail.

First version. Both interfering (compared) waves are reconstructed from a hologram on which the object was recorded twice (*double-exposure interferometry*). The pattern of fringes on the object's image represents the changes that occurred on the surface during the interval between two exposures. For this reason the technique is also referred to as *time-lapse* (or *lapsed-time*) *interferometry*.

It is extremely important that the position of the object during the second exposure be absolutely identical to that during the first exposure. This enables the two images recorded from the hologram to be precisely "inserted" one into the other.

The double-exposure technique is illustrated in Fig. 40. Figure 40*a* shows the hologram recording stage. The object's deformation is grossly exaggerated for the sake of clarity. The process of reconstruction is shown in Fig. 40*b*.

Second version. One of the interfering (compared) light waves is reconstructed from a hologram that was exposed only once. The second wave is scattered by the object itself, which is present when the hologram is being reconstructed. In order to achieve interference, the object must be precisely "inserted" into the virtual image reconstructed by the hologram. This requires that the object remain fixed on the bench during the interval between the hologram recording and the observation of the interferogram. The same requirement

applies to the hologram, so that it must be either processed on the spot or its material must be such that special processing is unnecessary.

The interferogram obtained with this technique enables us to estimate the changes on the surface of the object that have occurred by the moment of

Fig. 40

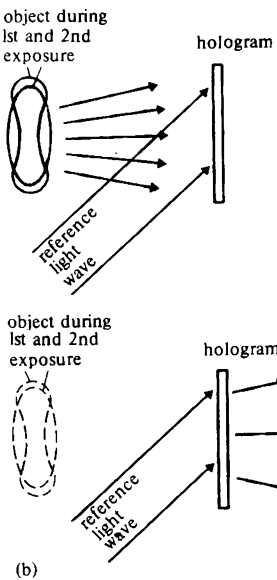
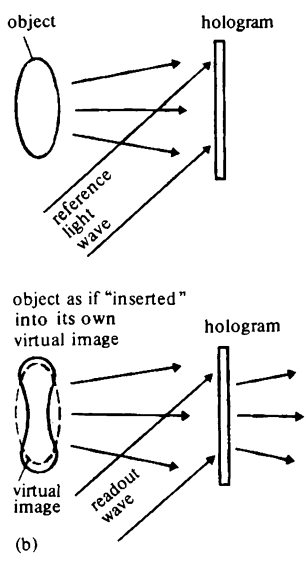


Fig. 41



observation. Furthermore, these changes can be followed as functions of time: the interference fringes may be modified in the process of observation. This technique is called therefore *real-time interferometry* (or *livefringe interferometry*).

The arrangement for real-time interferometry is illustrated in Fig. 41. The process of recording is shown in Fig. 41*a*; the reconstruction is illustrated in Fig. 41*b*, that is, observation of "live" fringes.

A comparison of Figs. 40 and 41 helps in clarifying the difference between the double-exposure and real-time methods of holographic interferometry.

Practical applications of holographic interferometry. The progress of technology demands ever increasing precision in manufacturing various products. This precision has to be verified. When this problem is tackled, any dismantling (in other words, destruction) of a unit is often out of the question, and even handling of the unit by "feelers", gauges, and so on is absolutely forbidden. Consequently, the problem has to be solved by non-destructive, contactless means of inspection. Holographic interferometry offers one of such means.

Assume that an object is placed in working conditions and undergoes various mechanical loads. We need to know what the distribution of internal stress in the object is, what regions are stress-concentrators and are therefore fraught with danger of failure. Holographic interferometry enables us to determine the degree and nature of deformation of the observed surface, that is, to extract the data necessary for calculating the internal stress distribution.

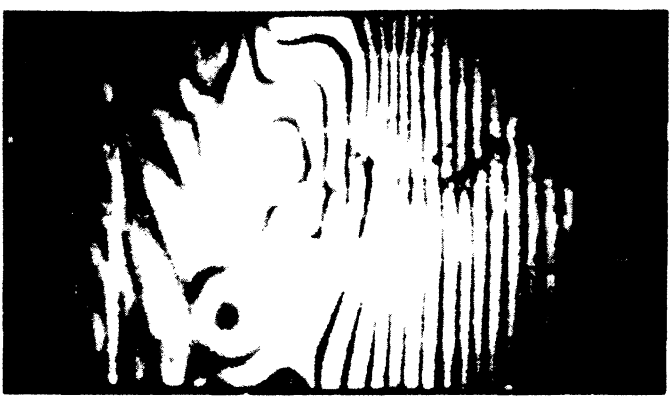
Obviously, internal stress induced in the object by changes in temperature can be analyzed in the same

manner. For instance, this is important for inspection of welded seams of metals with different coefficients of thermal expansion.

An object may have hidden internal defects (cracks, cavities, unwelded internal joints, etc.). In the case of metal objects, X-ray transmission methods are helpless. In some situations acoustic techniques fail also. Holographic interferometry may be applied to stressed objects, so that the interferogram shows the distribution of deformations over the surface, from which it is possible to establish the presence of defects, to determine their types, and even to locate them.

As an example, let us consider an object consisting of two welded metal plates; unwelded areas are possible on the internal surface of the seam. The object is statically strained. An interferogram (Fig. 42) clearly shows the regions where the regularity in the

Fig. 42

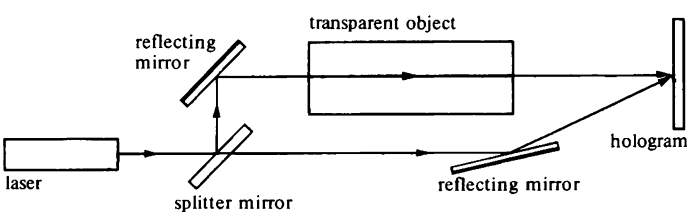


interference pattern is interrupted, which indicates the presence of internal defects in these places.

Holographic interferometry is also used to monitor the dimensions and shape of objects, and the quality of machining of their surface. For this purpose one of the compared waves is scattered by the object to be analyzed and the other is reconstructed by a hologram on which a reference object was previously recorded.

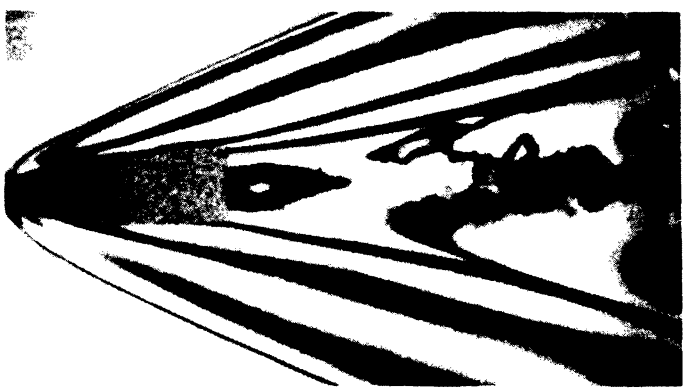
Holographic interferometry of transparent objects. The arrangement for recording holograms of transparent objects is shown in Fig. 43. The diagram does not show micro-objectives and individual lenses used to expand the light beams. When a light wave passes through a transparent object, it is modulated by specific phase information; the object modifies the pathlength of optical waves (the pathlength depends on the refractive index, and therefore on the density of the material) and, as a result, the phase of the waves. For this reason transparent objects are sometimes referred to as phase objects, and holograms of such objects — as *phase holograms.*

Fig. 43



Consider an example of holographic interferometry of a transparent object by means of the double-exposure technique. Let this object be a gas-filled chamber that is traversed by a bullet. The first exposure is taken before the shot, and the second during the passage of the bullet through the chamber (the laser pulse must be automatically triggered by the bullet's penetration into the chamber). Two light waves are reconstructed by the hologram; in one of them phase changes are determined by the gas in the unperturbed chamber, and in the other by the shock waves generated by the bullet. Note that shock waves change the gas density, and therefore the optical pathlength. Interference of light waves produces an interferogram whose structure characterizes density changes in the gas between exposures. The interferogram obtained is presented in Fig. 44.

Fig. 44



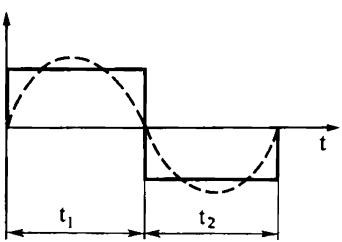
To summarize, holographic interferometry makes it possible to obtain instantaneous and very informative images of bulk distribution of density in gases, that is, to study in detail a number of gas dynamics phenomena.

Holographic interferometry of vibrating objects. Holographic interferometry is also widely used to analyze vibration of the surface of a number of objects.

In principle, an interferogram of a vibrating surface is very easy to obtain. It is sufficient to expose the film for a time much longer than the vibration period. The resultant hologram mostly reconstructs the two light waves that correspond to the two extreme positions of the oscillating portions of the surface. Interference of these two waves produces an interferogram typical of the given oscillating surface (so-called time-averaged interferometry).

In order to clarify these statements, let us consider Fig. 45, in which a sine wave characterizing the oscillatory motion is approximated by a step function. If this step curve really described the process of

Fig. 45

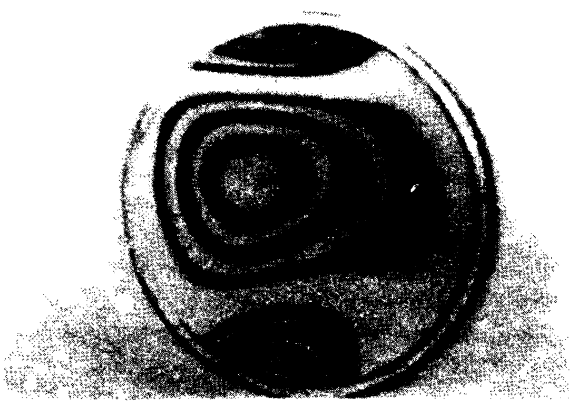


vibration, the hologram would indeed reconstruct only two object waves, one corresponding to the exposure during the interval t_1 and the other to the exposure during t_2. In fact, this would be identical to the double-exposure method. Obviously, this must in general remain correct when the step function is replaced by a sine wave.

An interferogram of the vibrating surface of a circular membrane is shown as an example in Fig. 46.

It should be noted in conclusion that the examples given above point conclusively to the practical significance of holographic interferometry. This technique is successfully applied to nondestructive control and inspection of both small and large objects (from miniature electronic devices and their assemblies to build-

Fig. 46



ings), as well as to detailed investigations of processes in gaseous mediums. In many cases the results obtained by means of holographic interferometry are unique.

11. Computer Technology and Holography

The links between computer science and holography are now well established and are developing fruitfully. This will be illustrated by a number of examples.

Computer-generated holograms. It has been said already that a hologram is, as a rule, a "pattern" on a plane. This is indeed an interference pattern formed in the hologram plane as a result of superposition of the object and reference waves. Although this pattern is complicated and involves minute details, in principle it can be reproduced by artificial means. Such synthetic holograms are now generated by computers.

The process of synthesis of a computer-generated hologram consists of several stages. First, the information on the shape of the surface of the object, whose hologram must be synthesized, is fed into a computer. The computer then calculates the amplitude and phase (i.e. the wavefront) of the object wave, that is, the wave that would be scattered by the object. To be precise, the computer calculates the amplitude and phase of the object wave in the plane of the future hologram. As the next step, the computer superposes the "object" wave and the "reference" wave, that is, computes the intensity distribution in the plane of the hologram. Finally, the computer transmits this pattern to a display device (such as a cathode-ray

tube), which graphically reproduces the obtained intensity distribution. The obtained "pattern" is then photographed. The developed negative represents a synthetic hologram.

At least two aspects make computer-generated holograms extremely interesting.

First, such holograms enable us to obtain visual 3-D "reconstructions" of imagined objects. For instance, one can reconstruct in three dimensions a "model" of an object still at the design stage. Such a model may eliminate the necessity of making an actual (material) model. One may visually "reconstruct" an object that never existed in real conditions.

Second, computer-generated holograms can be used to reconstruct light waves with a specified wavefront. This means that a specially computed and manufactured hologram may function as an optical element that transforms in a predetermined manner the incident (readout) light wave. The simplest example is the Fresnel zone-plate hologram, which is known to constitute a plana equivalent of the lens. It should be very attractive to substitute a set of optical elements (lenses, diaphragms, diffraction gratings, and so on) arranged in a specific manner by a single hologram

Calculation and manufacture of synthesized holograms form a new branch in optical holography, referred to as digital holography. In the general sense, digital holography comprises an analysis and synthesis of coherent optical fields by means of computers. No other field of holographic applications can use to such an extent the potentialities of controlling light waves.

Holography helps to manufacture microcircuits. The first digital computers appeared at the very beginning

of the 1950's. These were so-called first-generation computers. Their logical elements were electron tubes. The speed of operation of the first generation of computers was relatively low, from 10^2 to 10^4 operations per second. The main memory capacity of these computers was on the order of 10^4 bits*). Computers of the first generation were used for computations in problems of a purely scientific or commercial nature.

At the end of the 50's the second generation of computers was born. This generation was based on semiconductor triodes (transistors) used as logic elements. The rate of operation climbed to about 10^5 operations per second, and memory capacity increased to about 10^5 bits. Computers started processing large data arrays.

The third generation of computers, developed in the middle of the 60's, brought about a new era in the computer world: the era of data processing. It became possible to use computers to represent data in the informative format, to retrieve the required information from large data arrays, and to process this information. New computers can be given the task of controlling various processes. Third-generation computers have a very high speed of operation, up to 10^7 operations per second, and memory capacity up to 10^6 bits. The logic elements of these computers are based on semiconductor integral circuits (so-called *microcircuits*).

Microcircuits are also widely used in later generations of computers, and not only as logic

*) 1 bit is an elementary unit of information (the information sufficient to select one of two equiprobable events). Note for comparison that one volume of the Greater Soviet Encyclopedia contains about 10^7 bits of information.

elements, but also as memory devices. Microelectronics has become a decisive component of modern computer technology.

This brings to the fore the technological aspect of manufacturing microcircuits. A film microcircuit is a semiconductor film doped on predetermined areas by appropriate impurity atoms. These areas form an intricate and very detailed "pattern" on the surface, with some of the details being only 10 μm in size. A microcircuit is produced by using a special mask, that is, a very thin metal plate with holes reproducing the mentioned "pattern" (impurities are introduced into the semiconductor film through the holes in the mask). One of the techniques of preparation of such masks is based on photolithography.

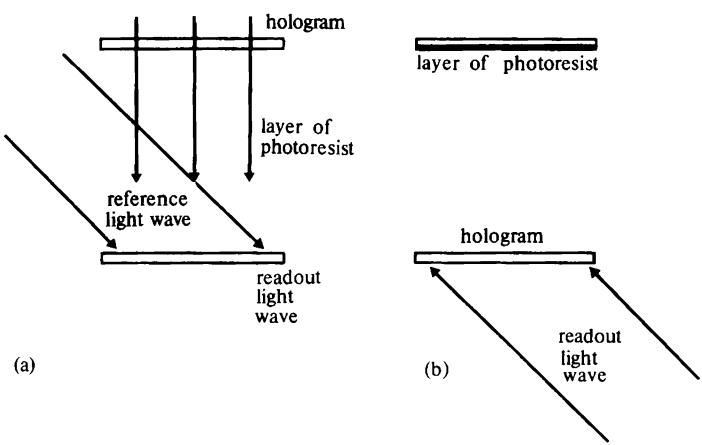
The diagram of a future microcircuit is first magnified to a scale of 10 : 1 or more. This magnified diagram is photographed with high resolution, and then printed at a scale of 1 1, which gives the so-called integrated circuit photomask. The mask itself is a copy of the photomask. For the copying, a thin metal plate is coated with a special light-sensitive lacquer (photoresist). If the photomask is placed over the photoresist and illuminated, the open areas of the photoresist layer are exposed in accordance with the configuration of the pattern on the photomask. This is the so-called contact method; another technique of transferring the pattern from the photomask to the photoresist layer, the so-called projection method, consists in the optical projection of the photomask diagram onto the photoresist.

Unexposed areas on the photoresist are then washed out, and the photoresist-coated metal film undergoes chemical etching. The etchant removes only those areas

on the film that are not protected by the photoresist layer, so that the result is the mask necessary for production of microcircuits.

Progress in computer technology results in constantly increasing requirements for microcircuits: higher area and finer details of the "pattern" At present large integral circuits require microcircuits with a pattern over an area with diameter of about 5 cm and with resolution of the order of 1 μm. High resolution over a large area is provided by the contact method of photomask application. Unfortunately, this method has certain shortcomings. The necessity of close contact of the photomask and photoresist layers results in wear and tear of the former. In addition, the method is rather difficult to automate.

Fig. 47



This aspect makes the method of holography very promising. Together with the advantages of the projection method (in fact, holography is a variety of this technique), holography produces high resolution (up to 1 μm) over a large area (up to 100 cm^2).

A holographic setup for printing the microcircuit diagram on the photoresist layer consists of two sections. First of all, the hologram of the photomask is recorded (see Fig. 47*a*). Then the hologram, after proper processing operations, is reconstructed in an arrangement that reconstructs an undistorted real image. To achieve this, the readout wave is directed on the hologram in the direction opposite to that of the reference wave (see Fig. 47*b*).

In addition to high resolution over a large area, the holographic technique of microcircuit exposure has a number of advantages directly related to the interference nature of the method. For instance, we could mention low sensitivity of photomask holograms to damage, the possibility of having a set of photomasks on a single hologram, the possibility of recording on curvilinear surfaces (by forming a 3-D image), and so on.

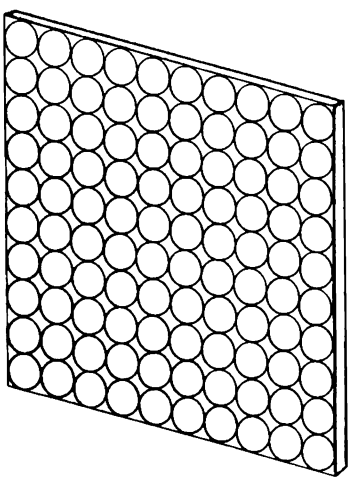
The holographic memory matrix: the key element of a holographic data storage device. At present the computer systems are more and more pervaded by optical methods of data processing. A trend is descernible of gradual "transformation" of digital computers into computer devices of a quite novel class, namely into optoelectronic systems. Data processing and data storage will be realized in these systems by means of both electric and coherent optical signals.

The first steps toward optoelectronic computers are

already taken. An example is found in computers that use an optical memory comprising holographic elements: computers with holographic data storage. Such systems are expected to have a great future because of the high capacity, reliability of storage, and high speed of data retrieval from the memory. Furthermore, holographic data storage seems to be the most efficient way of realizing associative memory in computers.

Holographic data storage is based on employing a holographic memory matrix. This matrix is composed of an array of small holograms 2 to 5 mm in diameter (see Fig. 48). Each of the holograms may carry from 10^4 to 10^6 bits of information. Each mini-hologram

Fig. 48

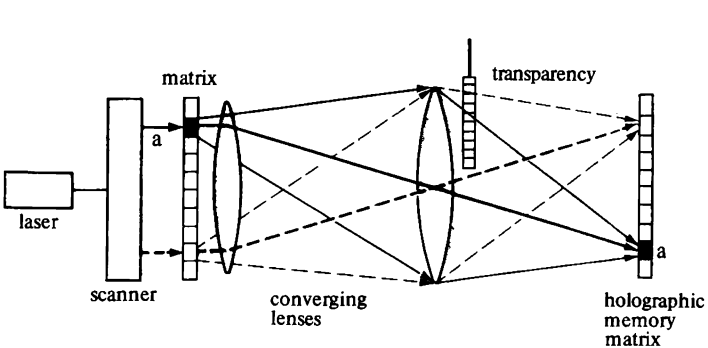


stores an interference pattern that is formed as a result of interference of the reference wave and a wave modulated by some signal. In the case of associative memory, the interference pattern results from superposition of the light wave modulated by the data stored and that modulated by the identification signal (associative signal).

Structure of the holographic data storage. Figure 49 shows one possible arrangement for information recording in the holographic memory.

One possible system of a transparency whose transmission is controlled by the external electric field is a matrix of liquid-crystal cells. Liquid crystals are specific organic dielectrics that possess, in a certain temperature range, properties intermediate between those of crystals and ordinary liquids. If voltage is applied to a cell of the transparency, a peculiar hydrodynamic effect is produced in this cell (in the liquid

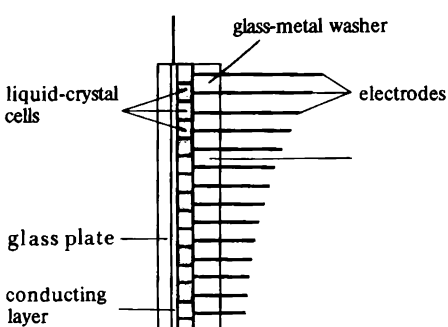
Fig. 49



crystal): interaction of the external field and the electric charges formed in the bulk of the liquid crystal produces turbulent macroscopic motions. This turbulence results in light scattering and, hence, in cell opacity. Thus, a layer of liquid crystal only 10 μm thick attenuates the light intensity by a factor of more than 10 at a voltage of 10 V. The cross-section of a liquid-crystal transparency is illustrated in Fig. 50.

Let us return to the diagram of the holographic data storage device shown in Fig. 49. Assume that the scanner directs the laser beam to hologram "a" in matrix (see the figure). The readout of this hologram reconstructs a divergent light beam (indicated in the figure by curvilinear hatching). At the same time, the hologram transmits the undisturbed laser beam, which is sent by the left-hand lens through the centre of the right-hand lens and onto the hologram "a" of the memory matrix (this beam is traced in the figure by a double line). The divergent light wave is spatially

Fig. 50



modulated by transparency and is also incident on hologram "a" of matrix. The interference of this wave with the light beam drawn by the double line records the information channeled to transparency into hologram "a" of matrix.

Let us change the information sent to transparency (by changing the set of electric signals applied to the transparency cells) and at the same time send the laser beam by means of scanner to a different hologram in matrix (see dashed lines in Fig. 49). Clearly, this new information is recorded on a different hologram in matrix. If scanner is synchronized with the device sending electric signals to transparency, the whole memory matrix can be gradually filled with the necessary information.

The above scheme was given as an example. It contains, however, all the basic elements involved in any holographic data storage device: a laser, a scanner, a transparency (we can term it the data input matrix), and a holographic memory matrix.

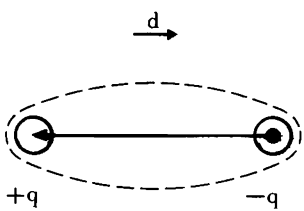
The examples discussed in this section sufficiently demonstrate the diverse relationships between the computer science and optical holography. We see that, on one hand, progress in computers helps the development of holography (the example of computer-generated holograms) and, on the other hand, holography participates in bringing new advances in computer systems. The contribution of holography to computer circuitry is determined both by its participation in improving the technology of computer manufacturing (the example of microcircuit technology), and by the important role that holographic data storage is expected to play in the future of optoelectronic computer systems.

12. A Few Words About Optical Characteristics of the Medium

Polarization of dielectrics. A dielectric can be polarized if it is placed, for example, between charged plates of a capacitor. Let us recall what processes take place inside the dielectric.

Assume that the dielectric is composed of polar molecules representing electric dipoles. Each such dipole is characterized by a physical vector quantity known as the electric dipole moment. Figure 51 schematically shows a dipole molecule as a system of two point charges $+q$ and $-q$ at a distance d from each other. The vector $\vec{d}$ is directed from the negative charge to the positive counterpart. By definition, the electric

Fig. 51



dipole moment is the vector

$$\vec{p} = q\vec{d} \tag{12.1}$$

In the case of zero external electric field, dipole moments point in random directions (Fig. 52*a*).

An external electric field $\vec{E}$ is a factor that tends to orient the dipole moments in the direction of the field. This trend is counteracted, however, by the thermal motion of molecules. As a result, the external field achieves only a partial orientation of dipole moments, as shown schematically in Fig. 52*b*.

Let us denote the sum of all dipole moment vectors per unit volume of the dielectric by $\vec{P}$

$$\vec{P} = \sum_i \vec{p}_i \tag{12.2}$$

where the subscript i enumerates individual dipole moments. Obviously, this vector is zero in the case shown in Fig. 52*a*, and distinct from zero in the situation of Fig. 52*b* (its direction evidently coincides with that of $\vec{E}$). The magnitude of $\vec{P}$ must be the greater the stronger the orienting effect of the external field is. This means that $\vec{P}$ gives a quantitative measure of polarization of dielectrics. It is called the *polarization vector*.

The orienting effect of the external field on molecular dipoles depends both on the properties of a medium and on field strength. Let us write therefore

$$\vec{P} = \varkappa \vec{E} \tag{12.3}$$

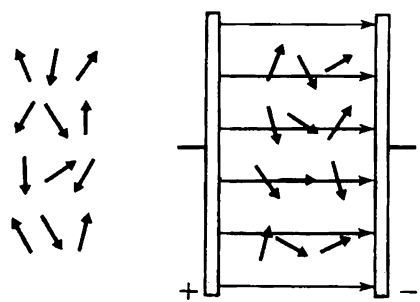
The factor $\varkappa$ characterizes the medium and is termed its *polarizability*, or *dielectric susceptibility*.

Note that Eq. (12.3) is an example of a material equation, that is of a relationship between an external factor and the "response" of the medium to this factor. In the caso under discussion, field strength $\vec{E}$ is the external factor and polarization vector $\vec{P}$ of the medium is its "response" to this factor.

Equation (12.3) is only one example of material equations. The other example, well known to the reader, is Ohm's law, in which the electric field strength (or the potential difference which determines the field strength) is an external factor and current density (or current) is the "response"

Light polarizes dielectric mediums. We assumed above that the dielectric was polarized in the electric field of a capacitor. This capacitor is in fact quite unnecessary. Dielectrics are polarized just as well by the electric field of a light wave propagating in them.

Fig. 52



In this case vector $\vec{E}$ in Eq. (12.3) is the vector of electric field strength of the light wave.

In what follows, it is this polarization that will be relevant in this context. The dielectric susceptibility $\varkappa$ will be regarded therefore as one of the optical characteristics of the medium.

Two points must be emphasized when polarization of the medium by light waves is discussed. *First*, the direction of oscillation of vector $\vec{E}$ must be adequately regulated, that is, the light itself must be polarized. *Second*, the light wave electric field, in contrast to that of a capacitor, varies with time, and at very high frequency at that. It becomes necessary, therefore, to take into account a sort of electric "inertia" of the medium, when response $\vec{P}$ of the medium follows $\vec{E}$ with some time lag. In rigorous terms, polarization $\vec{P}$ at a given moment is determined by field $\vec{E}$ at the preceding moments of time.

Optical characteristics of the medium. Two more optical characteristics of the medium will be considered in addition to the dielectric susceptibility $\varkappa$: dielectric permittivity ε and absolute refractive index n. The three optical characteristics are coupled by the following relationships:

$$n = \sqrt{\varepsilon} \tag{12.4}$$

$$\varepsilon = 1 + 4\pi\varkappa \tag{12.5}$$

Note that interrelationship between the refractive index of the medium, its dielectric permittivity, and dielectric susceptibility gives an additional argument in

favour of the electromagnetic nature of optical phenomena.

What factors determine the refractive index? On one hand, the refractive index is a function of properties of the refractive medium (its chemical composition, phase composition, and temperature). On the other hand, it is determined by the characteristics of the radiation:

frequency of the light wave (dispersion of light),

and polarization of light.

Moreover, the refractive index in crystals may depend on the direction of propagation of the light wave.

Let us analyze in more detail the refractive index of crystals as a function of polarization and direction of propagation of light waves. This will take us into the field of *crystal optics.*

Crystal optics. Crystal optics studies optical properties of crystals. Crystals are known to be anisotropic, which means that their physical properties, and optical properties among them, are functions of direction inside the crystal. The dependencies are found to be essentially different in different types of crystals. Let us choose as an example a fairly widespread type of crystals, which is referred to in crystal optics as *uniaxial.*

Each uniaxial crystal has an inherent direction termed its *principal axis.* Optical properties of such crystals remain unaltered if the direction in which they are measured is rotated around the principal axis (Fig. 53). Any other change in the direction will change the properties. For instance, the refractive index will be

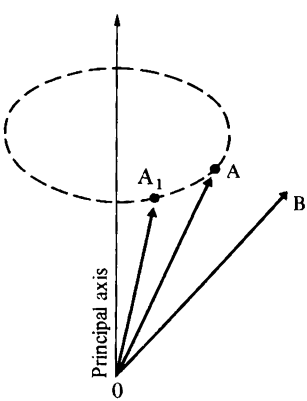
identical for directions OA and OA_1, and will differ for directions OA and OB.

The discussion to follow will concern only uniaxial crystals, owing to the predominant position they take in nonlinear optics.

Let us analyze propagation of light in a uniaxial crystal. The difference with an isotropic medium consists in the fact that two monochromatic waves, one "*ordinary*" and the other "*extraordinary*", both plane-polarized and having the same frequency, propagate simultaneously in any direction in a uniaxial crystal. It is of principal significance that each of these two waves has its own propagation velocity and, as a result, its own refractive index.

What is the difference between the ordinary and extraordinary waves? *First*, the plane of polarization

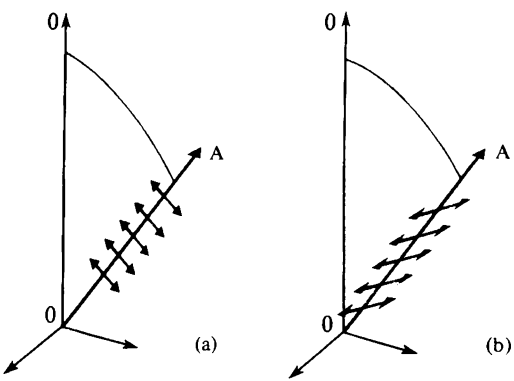
Fig. 53



(the plane in which oscillates the electric field vector) of the extraordinary wave lies in a plane drawn through the principal axis of the crystal and the direction of wave propagation (so-called *cardinal section plane*), while the plane of polarization of the ordinary wave is perpendicular to this plane. The situation is clarified in Fig. 54 which shows separately the cases of ordinary (*b*) and extraordinary (*a*) waves. Arrows in the figure indicate the direction of oscillation of the electric vector of the light wave; *OO* is the principal axis; *OA* is the chosen direction of wave propagation; the cardinal section plane is hatched.

Second, the refractive index of the ordinary wave (denoted by n_0) is independent of direction, while that of the extraordinary wave (denoted by n_e) is a function

Fig. 54



of direction:

$$\frac{1}{n_e} = \sqrt{\frac{\cos^2 \alpha}{n_1^2} + \frac{\sin^2 \alpha}{n_2^2}} \tag{12.6}$$

where α is an angle between the principal axis in the crystal and the wave propagation direction; n_1 and n_2 are optical parameters of a given uniaxial crystal called its principal indexes of refraction.

Note that the above comment on the difference between the indexes of refraction for the directions OA and OB (Fig. 53) was meant at the extraordinary wave: it is obvious that directions OA and OB differ in the value of α.

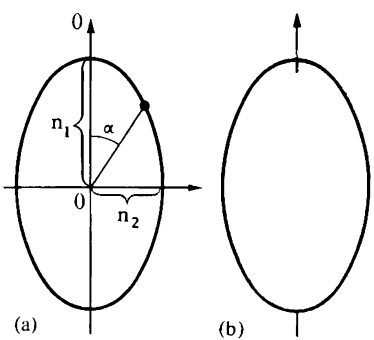
Refractive index as a function of α. Expression (12.6) gives the refractive index of the extraordinary wave, n_e, for any value of angle α. Let us draw a line OO on a sheet of paper, and assume it to be the principal axis of a crystal. Now draw a segment OA at an angle α to OO; the length of OA (in arbitrary units) is equal to the refractive index n_e for the same angle α. Now vary α from 0 to 360° and correspondingly rotate OA; the length of OA must be given by Eq. (12.6). The end point of OA (i.e. point A) will ultimately trace an ellipse (Fig. 55*a*). It is readily found that n_e at $\alpha = 0$ (and at $\alpha = 180°$) is equal to one of the principal refractive indexes ($n_e = n_1$), and at $\alpha = 90°$ (and $\alpha = 270°$) it is equal to other principal refractive index ($n_e = n_2$).

The obtained ellipse is a section of the *refractive index surface*, sometimes called the *indicatrix*, of the extraordinary wave. The indicatrix in Fig. 55*a* is drawn in the cardinal section plane.

The transition from the cardinal section plane, to the three-dimensional space is very simple since crystal rotation around the principal axis changes nothing in the physics of the phenomenon. Consequently, the refractive index surface of the extraordinary wave is the ellipsoid of revolution (Fig. 55*b*). If this ellipsoid is intersected by a plane passing through the principal axis of the crystal, we obtain the picture of Fig. 55*a*.

This analysis dealt with the refractive index surface of the extraordinary wave. As for the ordinary wave, its indicatrix is obviously spherical (circular in cross-section). It is important that the ordinary and extraordinary waves have identical refractive indexes in the direction of the principal axis; consequently, the radius of the refractive index surface of the ordinary wave is n_1. If we take account of both the extraordinary and ordinary waves, Figure 55*a* will be transformed to Figure 56 which shows, among other

Fig. 55



things, that the difference between refractive indexes of the ordinary and extraordinary waves reaches maximum at $\alpha = 90°$ (and $\alpha = 270°$).

The refractive index surface is an important optical characteristic of a crystal. It provides a simple method of finding the refractive index in any direction of wave propagation. For this, it is sufficient to draw a straight line from the indicatrix centre in the selected direction. The distance from the centre point to the intersection with the refractive index surface of the extraordinary wave yields the required value of n_e, and the distance to the surface corresponding to the ordinary wave is that of n_0.

The reader remembers that the extraordinary and ordinary waves are polarized in different planes (in mutually perpendicular planes). This means that Figure 56 also shows the refractive index as a function of polarization.

Fig. 56

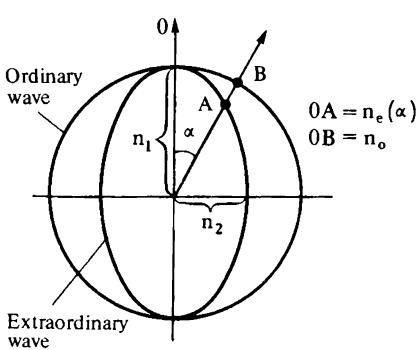


Fig. 57

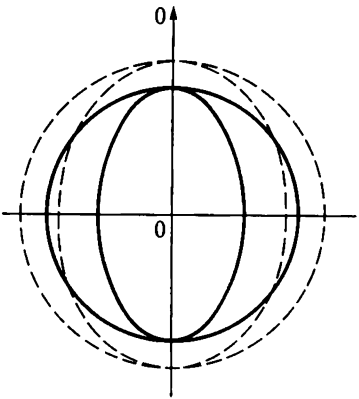
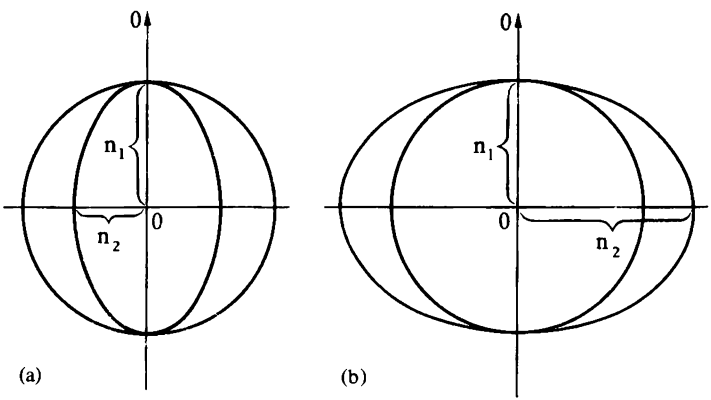


Fig. 58



Light dispersion affects the dimensions (but not the shape) of the refractive index surface; in other words, n_e and n_0 are functions of frequency. Sections drawn by dashed lines in Figure 57 correspond to higher frequency than those drawn by solid lines.

Figures 55–57 are drawn on the assumption that the principal refractive indexes of the crystal satisfy the inequality $n_1 > n_2$. It is customary to refer to uniaxial crystals of this type as *negative uniaxial crystals.* If, however, $n_1 < n_2$, a crystal is called the *positive uniaxial crystal.* Reference index surfaces of a negative (*a*) and positive (*b*) uniaxial crystals are compared in Figure 58.

13. Can the Optical Properties of a Medium Depend upon the Intensity of the Radiation?

The electric field of light waves is "gaining strength" Nothing was said until now about optical characteristics of the mediums as functions of the light wave intensity. Before the advent of the laser, incoherent optics correctly assumed that optical parameters of mediums are independent of the intensity of the light propagating in these mediums. The essential fact is that the electric field strength in fields emitted by non-laser light sources is always much smaller than field strengths of interatomic and atomic electric fields. Non-laser light sources generate fields with electric field strengths not exceeding 10^3 V/cm, while atomic fields are characterized by field strengths of the order of 10^7 to 10^{10} V/cm. It is only natural that with this ratio of field strengths, the light wave is

not intensive enough to affect atomic fields and with them the optical parameters.

Lasers have drastically changed the situation. Extremely high spatial concentration of light power became feasible owing to the high degree of coherence of the laser radiation. This is achieved, in practical terms, by very little divergence of the emitted beam and by the possibility to generate light pulses with very high peak power (see Sec. 4). Lasers made it possible to generate optical fields with field strength from 10^5 to 10^9 V/cm, which is already commensurate to that of atomic electric fields.

It has become necessary, therefore, to take account of the dependence of optical characteristics of the medium on the intensity of the light wave propagating in it. This necessity forms something of a "watershed" between the old (pre-laser) and new (laser) optics. It is customary to refer to the former as the *linear*, and to the latter as the *nonlinear optics*. It must be emphasized that dependence of optical parameters of the medium on the intensity of the light wave constitutes the most characteristic feature of nonlinear optics, one which distinguishes it from the linear optics.

What is the origin of the term "nonlinear optics" If the laser radiation is sufficiently intensive (so that optical characteristics are functions of light intensity), susceptibility $\varkappa$ stops being constant and becomes a *function* of field strength E in the light wave. Theory shows that in the first approximation this function can be expressed as a sum

$$\varkappa(E) = \varkappa_0 + \varkappa_1 E + \varkappa_2 E^2 + \ldots \qquad (13.1)$$

where $\varkappa_0$, $\varkappa_1$, $\varkappa_2$, are parameters of a medium characterizing its polarizability.

Note that all optical characteristics of mediums (not only susceptibility but dielectric permittivity and refractive index as well) become functions of field strength in sufficiently intensive light fields.

Substitution of Eq. (13.1) into (12.3) yields the following expression for polarization of the medium (vector notation is dropped for the sake of simplification):

$$P = \varkappa_0 E + \varkappa_1 E^2 + \varkappa_2 E^3 + \qquad (13.2)$$

An important fact is that Eq. (13.2) is *nonlinear* with respect to field strength in the light wave. Hence the term *nonlinear optics.*

If field strength in the light field is sufficiently low, only the first term can be retained in Eq. (13.2):

$$P = \varkappa_0 E \qquad (13.3)$$

This situation precisely corresponds to the pre-laser optics: polarization of the medium is described by a *linear* formula (13.3). Hence the term *linear optics.*

The relation between wave field strength and polarization of the medium is therefore linear if the light wave field strength is relatively low; the medium's polarizability is represented then by the parameter $\varkappa_0$ called the *linear susceptibility.* If, however, the light field strength in the laser beam is sufficiently high, the relation in question becomes nonlinear; additional parameters ($\varkappa_1$, $\varkappa_2$, ...) referred to as *nonlinear susceptibilities* are then required to describe polarizability of the medium.

Second-order and third-order nonlinear mediums. If polarization of a medium is described by a nonlinear expression, the medium is called a "nonlinear medium" Strictly speaking, any medium becomes "nonlinear" if a considerably intensive optical radiation is passed through it. The nonlinearity is found in the dependence of the medium's properties on light intensity.

Note that in optically isotropic mediums (such are not only gases and liquids but some crystals as well, namely those with cubic symmetry of the lattice) the expression for polarization does not comprise the quadratic term $\varkappa_1 E^2$ (owing to symmetry); the nonlinear relation (13.2) then becomes

$$P = \varkappa_0 E + \varkappa_2 E^3 \tag{13.4}$$

or, in vector notation,

$$\vec{P} = \varkappa_0 \vec{E} + \varkappa_2 E^2 \vec{E} \tag{13.5}$$

Both the quadratic and cubic "nonlinear terms" can be present only in the equation for optically anisotropic crystals, for instance for uniaxial crystals. We have to take into account, however, that as a rule the cubic (third-order) term is substantially smaller than the second-order one; hence, Eq. (13.2) is then simplified to *)

*) Note that Eq. (13.6) mathematically is not rigorous. Owing to anisotropy of the medium, vectors $\vec{P}$ and $\vec{E}$ are not parallel. Hence, we must consider not one but three simultaneous equations: each projection of vector $\vec{P}$ is expressed via three projections of vector $\vec{E}$ We consider that such a complication is out of place in a book like this, and so sacrifice some rigorousness and use mathematically simplified approximate expression of the type (13.2) and (13.6). This simplification is called *scalar*.

$$P = \varkappa_0 E + \varkappa_1 E^2 \tag{13.6}$$

We conclude, therefore, that the nonlinear expression for polarization of optically isotropic mediums will have, in addition to a linear term, only a third-order "nonlinear term" In this case the medium is said to have *third-order nonlinearity*. In the case of anisotropic crystals, Eq. (13.6) is valid, and the medium is said to have *second-order nonlinearity*.

What is the response of a nonlinear medium to external factors? We mentioned above that polarization P of a medium is a response of the medium to an external factor, namely to the field strength of the light wave. It is essential that the external factor under consideration is a function of time; for a monochromatic light wave, this function is

$$E = E_0 \cos(2\pi\nu t) \tag{13.7}$$

Our question now is: what will be the time dependence of the "response" of the medium?

In the case of a linear medium the "response" (i.e. polarization) will strictly follow the temporal changes in the external signal. Indeed, substitution of Eq. (13.7) into the linear relation (13.3) yields

$$P(t) = \varkappa_0 E_0 \cos(2\pi\nu t) \tag{13.8}$$

A nonlinear medium behaves in quite a different manner. The point to be emphasized here is that the response of a nonlinear medium is a different (sic!) function of time than that describing the applied external field. This is readily confirmed by substituting

Eq. (13.7) into a nonlinear relation (13.2). The substitution yields

$$P(t) = \varkappa_0 E_0 \cos(2\pi\nu t) + \varkappa_1 E_0^2 \cos^2(2\pi\nu t) + \\ + \varkappa_2 E_0^3 \cos^3(2\pi\nu t) + \tag{13.9}$$

If we use the formulas $\cos^2\alpha = (1 + \cos 2\alpha)/2$ and $\cos^3\alpha = (\cos 3\alpha + 3\cos\alpha)/4$, Eq (13.9) can be transformed to the form

$$P(t) = P_0 + P_1 \cos(2\pi\nu t) + P_2 \cos(4\pi\nu t) + \\ + P_3 \cos(6\pi\nu t) + \tag{13.10}$$

where

$$P_0 = \frac{1}{2}\varkappa_1 E_0^2 \tag{13.11}$$

$$P_1 = \varkappa_0 E_0 + \frac{3}{4}\varkappa_2 E_0^3 \tag{13.12}$$

$$P_2 = \frac{1}{2}\varkappa_1 E_0^2 \tag{13.13}$$

$$P_3 = \frac{1}{4}\varkappa_2 E_0^3 \tag{13.14}$$

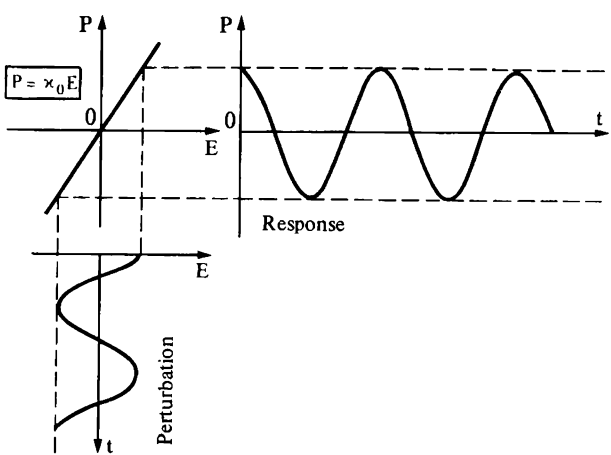
The response of the selected nonlinear medium to the external perturbation represented by a monochromatic light wave of frequency ν comprises four terms. The first one is independent of time; the second one follows the external perturbation (it is called the *first* or *fundamental harmonic of polarization*); the third one oscillates at frequency 2ν and is called

the *second harmonic of polarization*; the fourth term varies at frequency 3ν and is called the *third harmonic of polarization.*

The difference in the responses of the linear and nonlinear mediums to a monochromatic signal is clearly seen if one compares Figure 59 (the case of a linear medium) with Figure 60 (the case of a second-order-nonlinear medium). Figure 60 also demonstrates the above-mentioned components of polarization.

The third harmonic of polarization is absent if the medium is nonlinear to the second order. If the medium is third-order nonlinear, polarization contains no constant term and no second harmonic. In

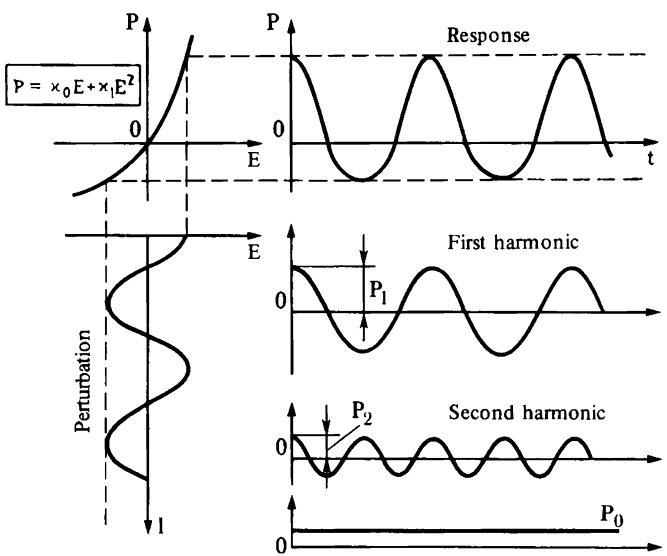
Fig. 59



a general case, the response of a nonlinear medium is determined by the form of function (13.1). Obviously, the response of the medium would be still more complicated if this function included terms with higher power of field strength E.

Nonlinear optical phenomena come to the fore. We should not forget, speaking about the time-dependent external factor (perturbation), that this factor is the light wave propagating in the medium at some velocity

Fig. 60



v. The polarization wave propagates in the medium at the same velocity. In linear mediums this wave has the same frequency ν as the light wave. In nonlinear mediums described by Eq. (13.2) several monochromatic polarization waves are produced having frequencies ν, 2ν, 3ν.

A polarization wave can be regarded as a sort of an "emitting antenna" moving at the velocity v through the medium. This "antenna" may emit a new light wave. Let us refer to this new light wave as re-emitted. The frequency of the re-emitted light wave must be equal to the polarization wave frequency; hence, nonlinear mediums may re-emit light waves not only at frequency ν but also at other frequencies, such as 2ν and 3ν.

Nonlinear polarization of the medium can thus lead to a specific nonlinear-optics phenomenon: transmission of a light wave with frequency ν is accompanied by emission of light waves at frequencies 2ν and 3ν. This phenomenon is referred to as *generation of optical harmonics.*

Generation of optical harmonics is only one of *phenomena of nonlinear optics* due to nonlinear polarization of the medium. Later we shall discuss other phenomena in the domain of nonlinear optics, also caused by specifics of the polarization response of a nonlinear medium to light waves.

Nonlinear optical phenomena as reponses of the medium to light waves. As follows from some of the above remarks, all phenomena in nonlinear optics are caused, in the long run, by changes in optical properties of the medium induced by sufficiently powerful optical radiation. This change in the

properties of the medium can be treated as a "response" to the light wave.

"Responses" can be of different origin. Two types are usually distinguished: the polarization response already mentioned, and the so-called "level population response"

We have already indicated that the *polarization response* is caused by nonlinear polarization of the medium induced by the incident light wave. Its field reorients electric dipole moments and also creates induced dipole moments. The polarization response is comparatively fast: its "inertia" is characterized by a fairly short time interval, down to 10^{-13} s.

The "level population response" is absolutely different in nature. It is related to changes in the population of energy levels again induced by the light wave propagating through the medium. Since this type of response involves transition of a large number of particles from one level to another, the response of the medium is found to be comparatively slow: its "inertia" is characterized by response time above 10^{-8} s. Consequently, the "level population response" cannot, in contrast to the polarization response, follow the light wave field without a considerable time lag.

Thus, fast and slow "responses" of the medium to an external perturbation (light wave) are distinguished. The former is of polarization nature, and the latter is associated with changes in population of energy levels.

Each of the two types of response is responsible for a specific group of nonlinear-optical effects. Those related to polarization response were already mentioned. Among the effects based on the "level population response" are such nonlinear effects as

induced transparency and induced opacity. We choose these two to begin a more detailed discussion of effects in nonlinear optics.

14. Intensity-dependent Transparency of the Medium

Could the photoelectric effect be observed beyond the low-energy threshold? The physical essence of the photoelectric effect is quite clear. In a few words, the phenomenon consists in ejection of electrons from the matter by the action of light. Assume that a photon with energy $h\nu$ is absorbed by an electron inside the matter. If the photon energy exceeds the work A required to remove an electron from this material (so-called work function), then the electron in question can leave the material. Energy of the ejected electron is

$$\frac{mv^2}{2} = h\nu - A \tag{14.1}$$

The photoelectric effect is observed, therefore, if the condition

$$h\nu \geqslant A \tag{14.2}$$

is satisfied. The frequency $\nu_0 = A/h$ is the lowest frequency for which photoelectric effect is observable in the material. It is called the threshold frequency.

Let us formulate the following question: could the photoelectric effect be observed beyond the threshold frequency, that is for $\nu < \nu_0$?

The theory of the photoelectric effect always assumed that an electron can absorb one (only one!) photon. This assumption was in perfect agreement with the experiment in the pre-laser optics. The advent of the laser changed the situation drastically. Light beams (and light pulses) now available to the experimenter reached extremely high intensity. The density of photons in such beams (pulses) was so high that an electron could absorb simultaneously two (and even more!) photons. The well-known laws of the photoelectric effect had therefore to be reconsidered.

First, the concept of the threshold frequency lost its significance. If, for instance, an electron absorbs not one but N photons at a time, each having energy of $h\nu$, then Eq. (14.1) has to be replaced by the relation

$$\frac{mv^2}{2} = Nh\nu - A \tag{14.3}$$

This means that the photoelectric effect can be observed if

$$h\nu \geqslant \frac{A}{N} \tag{14.4}$$

We find that the threshold frequency $\nu_0 = A/h$ does not constitute the lower bound any more. In other words, ν_0 cannot be considered, as before, as the lowest frequency at which the effect is observable.

Second, the statement that the occurrence of photoelectric emission is totally determined by the frequency of the radiation and has nothing to do with its intensity, had to be discarded. It is apparent that

the higher the radiation intensity is, the greater the photon density; consequently, the greater the number of photons that can be absorbed by one electron. In view of this, it should be more logical to speak not about the disappearance of the threshold frequency but about its dependence on the intensity of the light wave. The threshold frequency is found as

$$\nu_0(N) = \frac{A}{Nh} \tag{14.5}$$

The higher the light intensity, the greater N is and hence, the lower the threshold frequency.

The coming of the laser thus opened a new page in the history of the photoelectric effect. It became possible to observe the so-called *multiquantum photoelectric effect.* This effect totally belongs to the realm of the nonlinear optics: it is essentially intensity-dependent.

Induced opacity of the medium. Both the single-quantum and multiquantum photoelectric effects discussed above represent the so-called *external* effect: absorption of light results in the emission of the electrons out of the material. Another possibility is the *internal* photoelectric effect, when the absorption of light transfers electrons from lower energy levels to upper ones. Both effects may be single- or multiquantum. The multiquantum internal photoelectric effect is at the basis of the *induced opacity,* one of the effects in nonlinear optics.

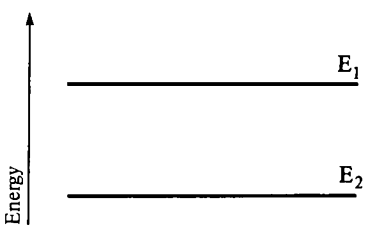
Let us consider the physical meaning of this phenomenon. We assume that light absorption occurs

because photons are adsorbed by certain particles of the medium (by absorption centres). Figure 61 shows energy levels of an absorption centre, where for the sake of simplicity only two levels, E_1 and E_2, are taken into consideration. Before irradiation, that is in the initial state, all absorption centres are at the ground level E_2. Assume now that the medium is irradiated with light whose frequency is chosen to be such that the photon energy be exactly one half of the difference between the energies of the two levels: $\nu = (E_1 - E_2)/2h$. A photon with energy $h\nu = (E_1 - E_2)/2$ cannot be absorbed by a centre since this energy is insufficient for transition from level E_2 to level E_1; hence, the medium is transparent to this radiation.

Now let us raise the intensity without changing frequency. Sufficiently high intensity provided only by the laser makes it possible to realize absorption of two photons simultaneously by a single absorption centre. This means that the centre gains energy $2h\nu = E_1 \quad E_2$ and jumps from level E_2 to level E_1. The medium is therefore capable of absorbing the optical radiation.

The simple example given above is a clear

Fig. 61



demonstration of the nonlinear phenomenon under consideration. Photons are not absorbed, and the medium remains transparent, when intensity is kept low. If intensity becomes sufficiently high, the photon density in the incident radiation increases to an extent where groups of photons begin to interact with absorption centres (groups in the above example are *pairs* of photons). Photon absorption is now possible and the medium ceases to be transparent.

Induced transparency of the medium. Not only opacity but transparency as well can be induced by light if its intensity is sufficiently high. In other words, an increase in light intensity which in some cases reduces transparency of the medium, in other cases may produce an opposite effect, that is *clarification of the medium.*

In essence, the physical meaning of this effect can be illustrated by using Fig. 61. But in this case we assume that the radiation frequency is $\nu = (E_1 - E_2)/h$. Photons with energy $h\nu = E_1 - E_2$ will of course be absorbed by the absorption centres. If the photon density in the radiation is sufficiently high, practically all absorption centres may be raised from level E_2 to level E_1. If this situation is realized, the medium is unable to absorb light at frequency $\nu = (E_1 - E_2)/h$: indeed, there is "nowhere" for the photons with energy $h\nu$ to be absorbed. Absorption of light at frequency ν reaches *saturation.*

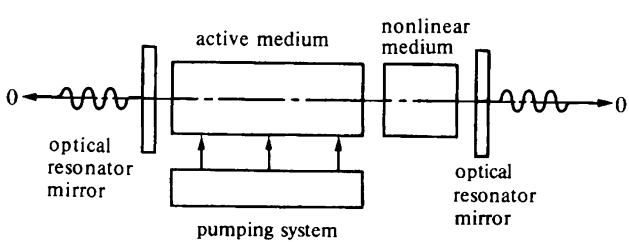
The effect of induced transparency is quite impressive: a powerful light pulse incident on an opaque medium makes it transparent almost instantaneously, and is transmitted. Some time after the end of the pulse (this time is different in different

mediums, and may vary in a rather wide range, from 10^{-8} to 10^{-2} s) the absorption centres return spontaneously to the ground level E_2 and the medium regains its opacity, at least until the arrival of the next powerful light pulse.

Phototropic gate: principle of functioning. The induced transparency effect is used in *phototropic gates* (shutters) already mentioned in connection with realization of the giant-pulse mode in lasers (see Sec. 4). A schematic of a laser with the phototropic gate is shown in Fig. 62 where *OO* is an optical axis of the laser.

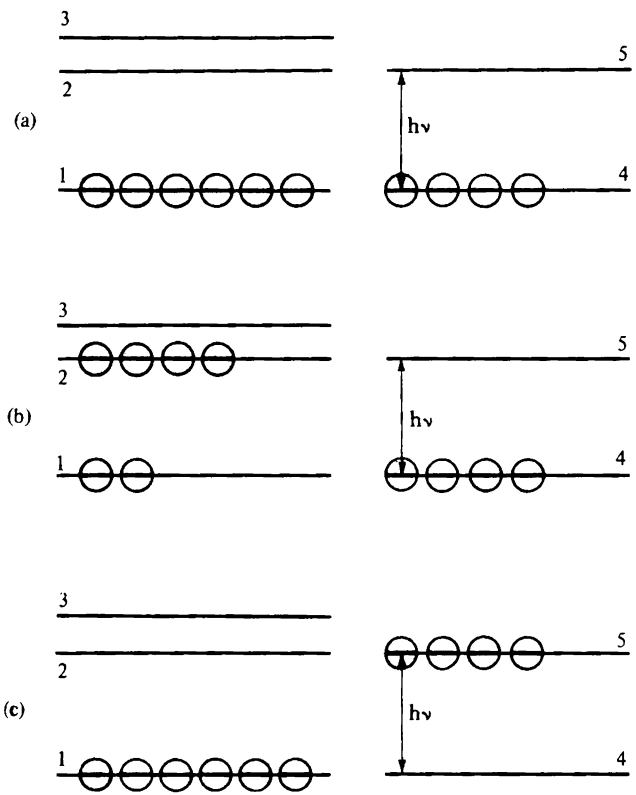
The process which results in emission of a giant pulse will be analyzed stage-by-stage, for reasons of convenience. The stages are illustrated in Fig. 63. Two systems of energy levels are shown for each stage: a system of three levels (levels *1*, *2*, *3*) of the active centre in the lasing medium, and a system of two levels (levels *4* and *5*) of the absorption centre in the induced-transparency medium, that is in the gate. It is essential that the difference between energy levels *4* and

Fig. 62



5 is equal to $h\nu$, that is to the energy of photons emitted by the active centres in transitions between lasing levels.

Fig. 63



The situation in Fig. 63*a* is the initial one (prior to the arrival of the pumping pulse): all active centres are at level *1*, all absorption centres at level *4*.

The situation of Fig. 63*b* occurs immediately after the pumping pulse: population of level *2* of the active centres is sufficiently high (population of the lasing levels is inverted). Note that the gate is opaque both in case (*a*) and in case (*b*); hence, attenuation for photon states with energy $h\nu$ is very high.

Situation (*b*) is unstable. Some of the active centres spontaneously return from level *2* to level *1*, emitting photons with energy $h\nu$. When these photons pass through the gate, they may be absorbed; as a result, absorption centres will start populating level *5*, thereby clarifying the gate medium. This induced transparency of the gate lowers the attenuation for photon states with energy $h\nu$, and at a certain moment the lasing condition will be satisfied. From this moment on, the process develops with accelerating pace. An avalanche of stimulated photons with energy $h\nu$ rapidly makes the gate completely transparent and, during its short life, generates a very narrow but tremendously powerful light pulse — the *giant pulse*.

Figure 63*c* illustrates the state of the system immediately after the emission of a giant pulse: all lasing centres are back at level *1*, and all absorption centres are at level *5* which corresponds to the maximum transparency of the gate.

Active centres will gradually return to level *4*. As a result, the gate will restore its opacity and photon states with energy $h\nu$ will be again effectively damped. In other words, both the lasing medium and the gate will be back at the initial configuration (*a*), ready to respond to another pumping pulse.

Compared to other types of optical gates (mechanical or electrooptical), phototropic gates have an important advantage, namely, they are automatic. The gate is opened by the pumping pulse and returns to the "initial position" automatically, immediately after the end of the giant pulse emission. The experimenter need not monitor, for instance, "on" and "off" states of electric or magnetic fields, or worry about synchronization of some rotating elements with pumping pulses, and so on. All this is made unnecessary. As to design considerations, a phototropic gate is simplicity itself, with all complexities "transferred" to the physics of the processes taking place in the nonlinear medium with induced transparency.

15. Self-focusing of Light

Refractive index as a function of intensity of light. Equations (12.4) and (12.5) cited above enable us to write

$$n = \sqrt{1 + 4\pi\varkappa} \tag{15.1}$$

Susceptibility $\varkappa$ of a nonlinear medium is a function of the light wave field strength; equation (15.1) then states that refractive index must also depend on this field strength. A nonlinear effect following from this is the *self-focusing of intensive light beams.*

Self-focusing does not alter the frequency of the light wave. This means that in analyzing the polarization response, we need to take account only of the term (13.12) which describes the fundamental

harmonic, that is the one with frequency equal to that of the light wave. By using Eqs. (13.10), (13.12), and (13.7), we can recast this harmonic in the form

$$P_{\nu}(t) = \left(\varkappa_0 + \frac{3}{4}\varkappa_2 E_0^2\right) E \tag{15.2}$$

This shows that instead of $\varkappa$, the expression for the refractive index (15.1) must include the sum $\varkappa_0 + \frac{3}{4}\varkappa_2 E_0^2$. Therefore,

$$n = \sqrt{1 + 4\pi\varkappa_0 + 3\pi\varkappa_2 E_0^2}$$

or

$$n = \sqrt{\varepsilon_l + \varepsilon_{nl}} \tag{15.3}$$

where $\varepsilon_l = 1 + 4\pi\varkappa_0$ is the dielectric permittivity of the corresponding linear medium, and $\varepsilon_{nl} = 3\pi\varkappa_2 E_0^2$ is a nonlinear increment in the expression for dielectric permittivity. Expression (15.3) can be simplified if we take into account that $\varepsilon_{nl} \ll \varepsilon_l$. By using an approximation formula $\sqrt{1+\beta} \approx 1 + \frac{1}{2}\beta$, valid for $\beta \ll 1$, we obtain from Eq. (15.3):

$$n = \sqrt{\varepsilon_l}\sqrt{1 + \frac{\varepsilon_{nl}}{\varepsilon_l}} \approx n_l(1 + \eta E_0^2) \tag{15.4}$$

Here $n_l = \sqrt{\varepsilon_l}$ is the refractive index of the linear medium, and $n_l \eta E_0^2 = \varepsilon_{nl}/(2\sqrt{\varepsilon_l})$ is the nonlinear increment in the expression for the refractive index. This increment is essentially proportional to the squared amplitude of the light wave, that is to its intensity.

Self-focusing. The refractive index of a nonlinear medium is therefore proportional to light intensity. Let us demonstrate that this results in self-focusing of light waves.

The intensity of any real light beam is not constant over its cross-section. Normally it peaks in the central area of the beam (in the vicinity of the optical axis of the beam) and falls off gradually away from the beam axis. This means, according to Eq. (15.4), that the refractive index of the medium (and so its optical thickness) must decrease away from the optical axis of the beam. Recalling Eq. (2.1), we conclude that the light propagation velocity must increase with the distance away from the beam axis.

Let us "picture" the light beam as a set of light rays, as is customary in the geometrical optics. The conclusion drawn above means that the rays farther from the beam axis will have greater velocity. Consequently, a plane wavefront incident on the material becomes concave in the process of wave propagation (Fig. 64). In other words, the beam transforms itself as if by a converging lens! It "contracts" toward the optical axis, in other words, it self-focuses.

Fig. 64

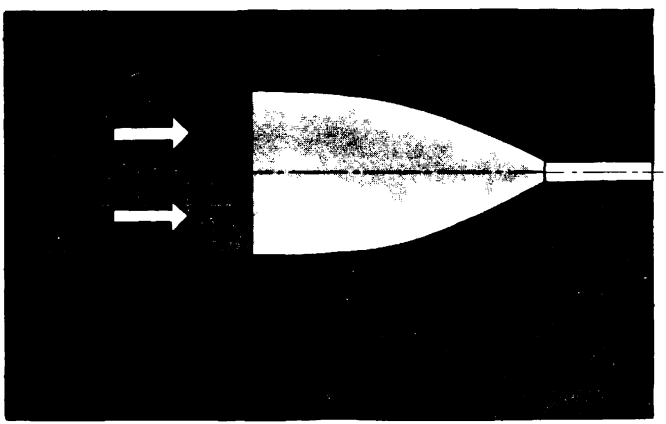
Self-focusing in practice. Figure 65 shows the way in which a light beam self-focuses in an appropriate medium. A light beam with diameter D, incident on a nonlinear medium, self-focuses over a distance L_0 and then propagates as a narrow light fibre. The distance L_0 can be estimated by an approximation formula

$$L_0 \approx \frac{D}{\sqrt{\eta E_0^2}}$$

The distance over which self-focusing develops is the shorter the higher the intensity and the greater the nonlinear susceptibility $\varkappa_2$ are.

Self-focusing is one of the threshold phenomena in nonlinear optics: it occurs when intensity reaches

Fig. 65



a certain limiting (threshold) value. The theory gives the following estimate for the threshold intensity:

$$I_{thr} = \frac{\lambda^2}{n_l^2 \eta D^2}$$

where λ is the light wavelength. The formula shows that the threshold intensity diminishes as the frequency of radiation rises. Furthermore, it is the lower the greater the nonlinear susceptibility $\varkappa_2$.

Self-focusing was mostly investigated in liquids: carbon disulphide, nitrobenzene, benzene, acetone, and some others. The observed light fibres were 30 to 50 μm in diameter, with the self-focusing distance L_0 being about 10 cm for the initial light-beam diameter of 0.5 mm.

Investigations show that light self-focusing is a very complicated phenomenon. It has been established, for example, that the observed light fibre has still finer structure – it separates into a number of still thinner "filaments" with diameters down to 5 μm.

16. Optical Transitions

Nonlinear optics proved successful in a solution of a problem which was very important both scientifically and practically: transformation of one light wave into another. For instance, how to transform a coherent light wave with frequency ν into a coherent light wave with frequency, say, 2ν?

Let us approach this problem first in the framework of *elementary processes*; in other words, in

terms of elementary interactions between individual photons and individual quantum systems.

There is no direct photon-to-photon interaction. We start with mentioning a fact of principal importance: there is no direct interaction between photons. True, we mentioned earlier the tendency of photons to populate first of all the states which already have sufficiently dense population. In a sense, this can be regarded as a sort of mutual "attraction" between photons. It should be kept in mind, however, that the tendency has nothing to do with the concept of "direct interaction" which causes scattering of particles on one another, absorption of some particles by other ones, and mutual transformations of particles, including decays. Photons are not scattered by photons, do not absorb one another, and do not decay. Neither electromagnetic nor other known forces mediate interactions between photons.

So there is no direct photon-to-photon interaction, and each time some photons are "transformed" into different ones, one must speak about photons interacting through a "go-between" The role of the intermediary is played by the matter, or to be exact, by its particles, first of all, by electrons. In what follows we consider this intermediary as a microscopic object characterized by a system of energy levels.

A photon and a microscopic object are in direct interaction. This means that the microscopic object can absorb or emit photons (or absorb and emit them simultaneously). In the process, it goes through quantum transitions between specific energy levels. Photons being essential participants of these transitions, the transitions are said to be *optical.*

All processes of "transformation" of photons into other photons (all processes of transformation of light into light) are therefore reduced to optical transitions in microscopic objects. For this reason, they deserve a more detailed discussion.

Single-photon and multi-photon transitions. Optical transitions are classified into *single-photon* and *multiphoton* ones. Only one photon is emitted or absorbed in a single-photon transition, while two or more participate simultaneously in multiphoton transitions. The number of photons participating in a transition determines its multiplicity: there are two-photon transitions (multiplicity *2*), three-photon transitions (multiplicity *3*), and so on. Let us consider the general case of a transition of multiplicity N. This means that N photons participate in this process. It may happen that m photons are emitted and $N - m$ absorbed. Obviously, all possible types of multiplicity N multiphoton transitions will be exhausted if m is varied from zero to N.

It should be emphasized that a multiphoton transition cannot in principle be divided into a temporal sequence of events; it must be treated as something indivisible in time.

As an example, let us take a two-photon transition in which two photons are absorbed. It would be wrong to assume that first one and then another photon is absorbed. It is essential that both photons are absorbed simultaneously. Otherwise we would have to consider not one two-photon transition but two single-photon transitions.

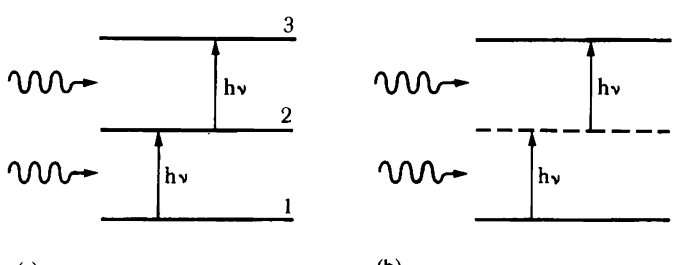
We conclude that any multiphoton transition is a qualitatively different process in comparison to a sequence (a set) of single-photon transitions.

What is a "virtual level"? Fig. 66*a* shows two single-photon transitions: first one photon with energy $h\nu$ is absorbed and the microscopic system is raised from level *1* to level *2*, and then the second photon is absorbed and the system is raised from level *2* to level *3*. But how do we symbolize a two-photon transition in which two photons of energy $h\nu$ are absorbed? By convention, it is shown as illustrated in Fig. 66*b*, where the dashed line denotes the so-called virtual level.

What is a "virtual level"? Recall, first of all, that a two-photon transition cannot be separated in time into two stages. Hence, the microscopic object cannot be observed on the virtual level (otherwise we could consider two stages — one prior to and another after the observation moment). This constitutes the main difference between a virtual and an ordinary energy level.

Could it be concluded then that a virtual level is "non-existent" or "not real"? Indeed, a microscopic system on any really existing level can always be observed, at least in principle.

Fig. 66

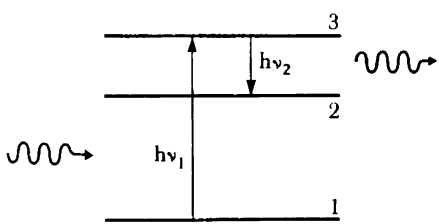


We are not going into a discussion of reality (or unreality) of virtual levels. The most important point for us is that both single- and multiphoton transitions are a reality. Moreover, a system of familiar (real) energy levels is sufficient to describe the behaviour of single-photon transitions, while in the case of multiphoton transitions this system is definitely insufficient and it is necessary to resort to a specific concept — that of a virtual level. The example in Fig. 66 is a clear illustration of the specific nature of this concept.

It should be noted in conclusion that one example of the two-photon absorption has already been discussed in Sec. 14 (the effect of induced opacity).

How the microscopic object acting as an intermediary in light-to-light transformation processes? Let us consider some of the processes in which photons are "transformed" into photons in a different state. We shall begin with a process shown in Fig. 67. The microscopic system absorbs a photon with energy $h\nu_1$ and changes from level *1* to level *3*. The system

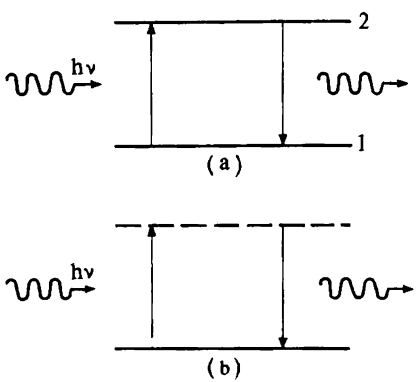
Fig. 67



then emits a photon with energy $h\nu_2$ and drops from level *3* to level *2*. The initial (primary) photon with energy $h\nu_1$ is therefore "transformed" into the final (secondary) photon with energy $h\nu_2$. The microscopic system has acted in this "transformation" as an "intermediary": indeed, its state has also been changed, from level *1* to level *2*.

This role of an intermediary between photons (but nothing more than an intermediary) stands out still more clearly in the process shown in Fig. 68*a*. The microscopic system absorbs a photon with energy $h\nu$ and changes from level *1* to level *2*. Then it emits a photon with the same energy and returns to level *1*. The state of the microscopic system finally remains unaltered, while the primary photon has been "turned" into the secondary one. The secondary photon has the

Fig. 68



same energy but of course may differ in both the momentum direction and polarization.

Let us turn now to the process shown in Fig. 68*b* (dashed line denotes a virtual level). In contrast to the two preceding single-photon processes, here we deal with one two-photon event rather than two single-photon events. In principle, the case of Fig. 68*a* allows to observe the microscopic system on level 2 (during the interval between the absorption of the primary photon and the emission of the secondary one). The situation of Fig. 68*b* is absolutely different: the microscopic system cannot be found on the virtual level in principle, since there is no "interval" between the absorption of the primary and the emission of the secondary photons. One cannot even state that the primary photon absorption occurs prior to the emission of the secondary one. The process of absorption and emission proceeds as something indivisible in time, so that it would be meaningless to try and observe even temporary changes in the state of the microscopic system.

The microscopic system of the above-discussed two-photon process may be said to behave as a very "tactful" intermediary staying very much "behind the scenes"

A process describing generation of the second harmonic (SHG). The multiphoton processes in which the initial and the final states of the microscopic system are identical, are of special interest in the nonlinear optics. The two-photon process was discussed above. Let us consider two *three-photon* processes.

The first of them is shown in Fig. 69 (dashed lines indicate virtual levels). The microscopic system

participates in a three-photon transition: two photons, with energy $h\nu$ each, are absorbed and one photon with energy $2h\nu$ is emitted; the state of the microscopic system is unaltered. The microscopic system as an "intermediary" remaining "in the shadows", we can regard the process as a "direct transformation" of two photons into one (two colliding photons merge into a single one). The energy-momentum conservation law is satisfied:

$$h\nu + h\nu = 2h\nu \tag{16.1}$$

$$\vec{p}_1 + \vec{p}_2 = \vec{p} \tag{16.1'}$$

(here $\vec{p_1}$ and $\vec{p_2}$ are the momenta of the absorbed photons, and $\vec{p}$ is the momentum of the emitted one).

In nonlinear optics the above process is referred to as the *second harmonic generation* (SHG). It describes the "transformation" of light with frequency ν into light with double frequency, 2ν. The SHG-process is treated in more detail in Sec. 17.

A process describing parametric generation of light. Figure 70 represents a three-photon process in which

Fig. 69

one photon with energy $h\nu$ is absorbed and two photons, one with energy $h\nu_1$ and the other with $h\nu_2$, are emitted; the state of the microscopic system is unaltered. In a sense, this process can be treated as a "decay" of one (primary) photon into two new (secondary) ones. The photons participating in the process are again subject to the energy-momentum conservation:

$$h\nu = h\nu_1 + h\nu_2 \tag{16.2}$$

$$\vec{p} = \vec{p}_1 + \vec{p}_2 \tag{16.2'}$$

This process is often referred to as the *parametric generation of light*. It describes the "transformation" of the light wave with frequency ν into two new light waves with frequencies ν_1 and ν_2. In principle, any of these frequencies (for instance, ν_1) can be varied in a continuous manner from zero to ν. The phenomenon of parametric generation of light will be analyzed in more detail in Sections 17 and 18.

On one possible source of doubt. A reader may be uncertain about whether the processes in Figs. 69 and

Fig. 70

70 indeed require to be mediated by an "intermediary" Can it be that the photons in these processes interact directly?

Indeed, we are tempted to assume that in some processes photons interact without intermediaries (as numerous other particles do). In this case we could get rid of the concept of a virtual level. It would be possible then to consider that a photon with energy $h\nu$ in Fig. 70 decays into two photons $h\nu_1$ and $h\nu_2$ without involving the microscopic system which remains at the same energy level and does not take part in virtual transitions.

These arguments are, however, unacceptable. Experiments show that the processes shown in Figs. 69 and 70 (and other processes in question) simply cannot occur outside the medium! However deep the "shadow" in which the microscopic system is lurking, its role of an intermediary is always decisive, in the sense that it determines the very possibility of realization of a multiphoton process.

17. Transformation of One Light Wave into the Other

Incoherent and coherent light-to-light transformation processes. The preceding subsection was devoted to various processes of light-to-light transformation illustrated by elementary interactions between photons and the microscopic system. Sometimes transitions with absorption of photons and those with emission of photons are sharply separated in time: they are accompanied by changes in the state of the microscopic system (even if the initial and final states of this system

are identical). In other processes these transitions are not separated in time, and no changes could be detected in the state of the microscopic system; the energy-momentum conservation in these processes holds as though photons interact directly.

The processes of the first type are referred to as the *incoherent* processes of light-to-light transformation, while the processes of the second type are said to be *coherent* processes. Let us analyze some specifics of both types of processes.

Incoherent processes. In the case of incoherent processes the primary light wave (pumping wave) is absorbed and thereby causes changes in the population of levels in the material. New quantum transitions in the medium then result in emission of the secondary light wave. Obviously, no interaction between the pumping wave and the secondary light wave is possible in this case. Indeed, first the pumping wave raises the medium to the excited state and only after some time the medium returns to the ground state and emits the secondary light wave.

The process of generation of the laser emission under conditions of optical pumping is one example of the incoherent light-to-light transformation. Here the radiation of the flash lamp is the pumping wave, and the coherent radiation generated in the lasant is the secondary light wave. Another example is the photoluminescence which is used in lamps customarily referred to as fluorescent lights.

Coherent processes. In contrast to incoherent processes, the acts of interaction of the medium with the pumping wave and with the secondary wave cannot be separated in time and have to be regarded as a unified process (we remind that this constitutes the

main distinction of transitions involving virtual levels). This characteristic of coherent processes is essential in two respects. *First*, it is impossible to detect any changes in the state of the medium interacting with light waves. *Second*, we can, in a sense, operate with direct interaction between the pumping wave and the secondary wave. No doubt, waves interact only via the matter and this interaction is determined by the matter's parameters. Nevertheless, the participation of the medium, though being essential in principle, remains *virtual in character* so that the light waves interact as if directly.

Interaction of the waves requires that the pumping and secondary waves *be matched* in frequency, direction of propagation, and polarization. This means that each of the interacting waves must obviously be characterized by a definite frequency, propagation direction, and polarization. Hence, coherent processes must involve highly coherent light waves. It can be said that all coherent processes are the processes of transformation of coherent light to coherent light.

The importance of light coherence to coherent processes can be additionally clarified on the basis of photon concepts. For a coherent process to occur, it is necessary that the momentum-energy conservation law be satisfied for the photons; consequently, the primary and secondary photons must be in definite states, that is in states with definite energy and momentum. Obviously, the greater the number of photons in the required state and the smaller the spread of photons over all other possible states, the more effective the coherent process in question is. And reduction of the spread of photons over allowed states is identical to increasing the coherence of the radiation (see Sec. 2).

The requirement of matching the pumping wave parameters with those of the secondary wave is formulated as the condition of *wave synchronism.* In "photon terms" this condition is identical to the law of momentum conservation for photons participating in the process. The wave synchronism condition is important in coherent processes: it is a necessary condition of the efficient transfer of light energy from the pumping wave to the secondary one.

Let us elucidate the meaning of the wave synchronism condition by analyzing an example of the second harmonic generation.

The condition of wave synchronism in the case of SHG. Assume that the directions of propagation of the pumping and the secondary wave in the case of SHG are identical, so that momentums of all photons in Eq. (16.1′) propagate in the same direction. This enables us to replace the vector equation (16.1′) by a scalar one:

$$2p_1 = p \tag{17.1}$$

where p_1 and p are momentums of the primary and secondary photons, respectively.

We remind that the photon momentum in vacuum is expressed in terms of the radiation frequency by means of Eq. (2.6). In the case of a medium this formula has to be slightly corrected by introducing the index of refraction (frequency-dependent):

$$p = \frac{h\nu}{c} n(\nu) \tag{17.2}$$

Making use of Eqs. (17.2) and (16.1), recast (17.1) to the form

$$2\frac{h\nu}{c}n(\nu) = \frac{h2\nu}{c}n(2\nu)$$

which yields, after factoring out identical multipliers,

$$n(\nu) = n(2\nu) \tag{17.3}$$

This is the wave synchronism condition for the SHG process. According to this condition, efficient transfer of light energy from the pumping wave to the second harmonic requires that the two waves have identical refractive indices.

Obviously, Eq. (17.3) does not hold in the general case (because of the dispersion effect). So the practically important problem becomes: how to satisfy this condition. A satisfactory answer to this question was not immediately found. It proved to be very interesting: it suggested using the dependence of the refractive index on the direction in the crystal.

Take a uniaxial crystal.... Refractive index surface of a negative uniaxial crystal was shown in Sec. 12 (see Fig. 57). These surfaces are repeated in Fig. 71, with solid curves tracing surfaces for frequency ν and dashed curves — for frequency 2ν. The refractive index surfaces of the ordinary wave with frequency ν and that of the extraordinary wave with frequency 2ν intersect in points A and A_1.

This means that if we select, for example, the direction AA (going at an angle α to the principal axis

of the crystal), then the following condition is satisfied for light waves propagating in the direction AA:

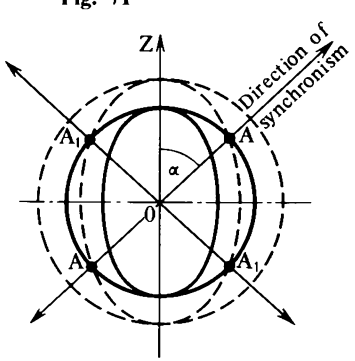
$$n_0(\nu) = n_e(2\nu) \tag{17.4}$$

This is the synchronism condition for the SHG process in which the pumping wave is the ordinary wave, and the second harmonic is the extraordinary wave. The direction AA is the *direction of synchronism* for the process in question.

What has to be done, therefore, in order to realize the process of generation of the second harmonic?

This requires first of all a uniaxial crystal with sufficiently high nonlinear susceptibility $\varkappa_1$. (This can be the negative uniaxial crystal of potassium dihydrophosphate, KH_2PO_4.) The crystal must be cut in the shape, for example, of a rectangular parallelepiped with the axis along the direction of

Fig. 71



synchronism for the given frequency ν of the pumping radiation.

Pumping must be realized with a laser. The pumping wave must be plane-polarized, with the plane of polarization perpendicular to the plane of cardinal section of the nonlinear crystal (the plane drawn through the principal axis of the crystal and the parallelepiped axis). This polarization of the pumping wave is required in order to act as the ordinary wave (the plane of polarization of the ordinary wave is perpendicular to the cardinal section plane).

If these conditions are satisfied, propagation of the pumping wave with frequency ν in a nonlinear crystal produces an additional light wave: the second optical harmonic. The direction of propagation of this wave coincides with that of the pumping wave (although the opposite direction is also possible), its frequency is twice that of the primary radiation, and the plane of polarization coincides with the cardinal section plane as in the extraordinary wave. If a nonlinear crystal is several centimetres long, more than 10% of the pumping energy can be transformed into the second harmonic radiation.

Classical interpretation of the second harmonic generation. Until now the SHG effect was described in photon terms, that is by referring to the three-photon process shown in Fig. 69. It is not difficult, however, to supply a purely classical explanation as well.

Let a coherent pumping wave (13.7) with frequency ν be incident on a quadratically nonlinear medium. If the medium were linear, its polarization would change in time exactly as the pumping wave field strength does i.e. with frequency ν [see Eq. (13.8)]. But

polarization of nonlinear mediums contains the second harmonic as well: the term $\frac{1}{2}\varkappa_1 E_0^2 \cos(4\pi\nu t)$ in Eq. (13.10). Naturally, polarization oscillating at frequency 2ν may result in re-emission of light at this double frequency, that is in the emission of the secondary light wave with frequency 2ν.

It has already been mentioned in Sec. 13 that the wave of polarization (and with it the second harmonic of polarization) propagates in the medium with the velocity of the pumping wave, that is with the velocity $c/n(\nu)$. Energy transfer from the polarization wave to the re-emitted light wave will only be efficient if wave propagation velocities coincide. The velocity of the wave with frequency 2ν being $c/n(2\nu)$, the light re-emission condition at this frequency becomes

$$n(\nu) = n(2\nu)$$

which is a familiar condition of wave synchronism.

This is a classical interpretation of the SHG effect in nonlinear optics. Note that this interpretation brings to the fore the role of the medium as an intermediary in the interaction between the primary and secondary light waves. Indeed, the interaction is "transferred" along a "chain": pumping wave — polarization wave — secondary light wave.

The process of the *third harmonic generation* can also be easily outlined. In "photon terms" this is a specific *four-photon* process in which three photons with energies $h\nu$ are annihilated and one photon with energy $3h\nu$ is produced. In terms of the classical wave concept this is a result of re-emission following directly from the existence of the third harmonic in the nonlinear polarization of the medium [see the term $\frac{1}{4}\varkappa_2 E_0^3 \cos(6\pi\nu t)$ in Eq. (13.10)].

The processes of generation of still higher orders of optical harmonics—fourth, fifth, and so on, are also possible. It seems that these processes do not follow from the formula (13.10). However, Eq. (13.10) was based on (13.1) which contains the terms of order not higher than E^2; higher powers of E were neglected. In principle, we could include into (13.1) several higher-order terms which would make it possible to analyze higher-order harmonics in the nonlinear polarization of the medium.

Nonlinear polarization of the medium allows mixing of frequencies. Let the polarization of a nonlinear medium be represented by Eq. (13.6). We assume that two coherent light waves with unequal frequencies are incident on the medium: $E_1 \cos(2\pi\nu_1 t)$ and $E_2 \cos(2\pi\nu_2 t)$. If the sum of these waves,

$$E = E_1 \cos(2\pi\nu_1 t) + E_2 \cos(2\pi\nu_2 t) \tag{17.5}$$

is substituted into Eq. (13.6), the final expression for the polarization of the medium will contain a term

$$P_{1,2} = 2\varkappa_1 E_1 E_2 \cos(2\pi\nu_1 t) \cos(2\pi\nu_2 t) \tag{17.6}$$

Making use of the relation $2\cos\alpha\cos\beta = \cos(\alpha + \beta) + \cos(\alpha - \beta)$, we transform Eq. (17.6) to the following:

$$P_{1,2} = \varkappa_1 E_1 E_2 \cos[2\pi(\nu_1 + \nu_2)t] + \\ + \varkappa_1 E_1 E_2 \cos[2\pi(\nu_1 - \nu_2)t] \tag{17.7}$$

The fact that the expression for nonlinear polarization of the medium contains the term (17.7) means that light can be re-emitted at frequencies $\nu_1 + \nu_2$ and $\nu_1 - \nu_2$. Hence, nonlinear mediums make it possible to realize *summation* and *subtraction* of light wave frequencies. We find that in the case in question interaction of waves with frequencies ν_1 and ν_2 may generate secondary light waves with frequencies $\nu_1 + \nu_2$ and $\nu_1 - \nu_2$.

Equation (13.6) is the simplest expression for the polarization of a nonlinear medium (nonlinear polarization is described by a term quadratic in field strength). In a more general case, polarization expression may include also the terms with E^3, E^4, and so on. With these terms, substitution into Eq. (13.5) results in polarization containing terms with frequencies $\nu_{mn} = n\nu_1 \pm m\nu_2$, where n and m are integers. This means that other types of *frequency mixing* are possible in addition to summation and subtraction.

We note in conclusion that the difference $\nu_1 - \nu_2$ may fall into the range of acoustic frequencies. In this sense one may speak of, for example, an optical method of generation of ultrasonic waves.

A classical explanation of the parametric generation of light. The effect of parametric generation of light was analyzed in Sec. 16 in terms of photon concepts. A classical explanation of this phenomenon is also possible. It is based on the discussed above "mixing" of light waves in a medium with nonlinear polarization.

Assume that the medium is illuminated simultaneously by a high-intensity coherent light wave (pumping wave) $E_p \cos(2\pi\nu t)$ and two low-intensity light waves as the first and second initial signals,

condition $\nu_1 + \nu_2 = \nu$. Let us refer to these weak waves as the first and second initial signals, respectively. "Mixing" of the first initial signal with the pumping wave may result in a sufficiently intensive secondary wave at frequency $\nu - \nu_1 = \nu_2$. Likewise, "mixing" of the second initial signal with the pumping wave may produce a secondary wave at frequency $\nu - \nu_2 = \nu_1$. We can therefore excite parametrically two secondary light waves at frequencies ν_1 and ν_2, at the expense of part of the energy of the pumping wave.

Naturally, parametric generation of light requires, as any other coherent process, that the corresponding condition of wave synchronism be satisfied. Materials to be used are, as in the case of SHG, uniaxial crystals with relatively high nonlinear susceptibility; moreover, the direction of synchronism is again found for the chosen combination of frequencies ν and ν_1.

Initial signals required for "triggering" the process of parametric generation are always available in any real crystal in the form of an inevitable "background" which can be explained, among other factors, by the presence of spontaneous photons. These very weak signals are "distributed" over a spectrum of photon states. The experimenter can, by varying the synchronism direction, bring "into play" initial signals of different frequencies and thus tune the frequencies ν_1 and ν_2 of the secondary light waves generated in the process, in a continuous manner.

The term *parametric generation of light* originates from the similarity between this process and the method of parametric excitation of oscillations widely used in electronics. Parametric phenomena in electronics occur in circuits involving nonlinear capacitors, while in optics we have to use nonlinear crystals.

18. The Principle of Operation of the Parametric Light Oscillator

Elements required to realize parametric generation of light. A general set of requirements was in fact given in the preceding Section.

The first element is a uniaxial crystal with relatively high nonlinear susceptibility, cut according to the requirements of the wave synchronism. This crystal is placed inside the optical resonator in order to direct the synchronism direction, for a given combination of frequencies ν_1 and $\nu_2 = \nu - \nu_1$, along the optical axis of the resonator. This will favour photon states with energies $h\nu_1$ and $h(\nu - \nu_1) = h\nu_2$. Now the resonator is illuminated by a coherent light wave with frequency ν (pumping wave emitted by a laser). The intensity of the pumping wave must be sufficiently high in order to bring out the nonlinear properties of the crystal and to exceed the level of losses for favoured photon states. Furthermore, in accordance with the requirements of the wave synchronism, the pumping wave must be either ordinary or extraordinary, that is it must be polarized in a specific manner. And finally, there must be a device which makes it possible to "control the direction of synchronism" thereby realizing smooth tunability of the frequency of secondary light waves.

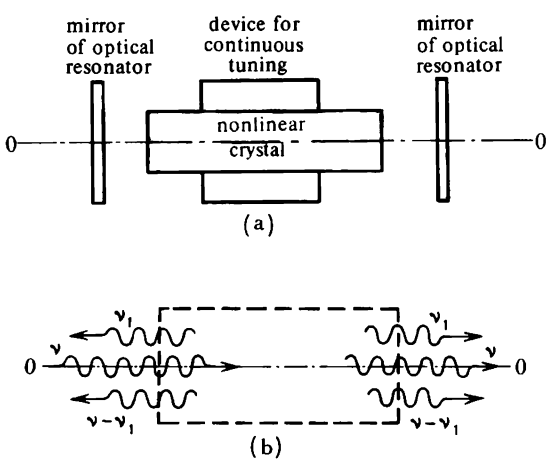
Parametric light oscillator. A schematic of the parametric light oscillator is shown in Fig. 72*a* (*OO* — optical axis of the system defined by the resonator mirrors).

Figure 72*b* schematically shows the pumping wave and secondary light waves generated in the parametric light oscillator which is shown by a rectangle. The

pumping wave with frequency ν propagates along the optical axis *OO*. Two secondary waves with frequencies ν_1 and $\nu_2 = \nu - \nu_1$ propagate along this axis in both directions. All interacting light waves in the configuration under discussion are directed along the optical axis *OO*; in other words, momentums of all photons involved in the process have identical direction.

The wave synchronism condition is satisfied, in a given crystal, for certain directions specific for each combination of frequencies ν_1 and $\nu_2 = \nu - \nu_1$; these directions (directions of synchronism) are at a certain angle to the principal axis of the crystal. Without going into the wave synchronism condition for the

Fig. 72



parametric generation of light (wave synchronism was analyzed in detail for the SHG case), let us note that this condition is satisfied only by special combinations of ordinary and extraordinary waves. Different combinations are possible. In *one* of them an extraordinary pumping wave is used; one of the secondary waves is extraordinary and the other is ordinary. In *another* combination, an ordinary pumping wave is used, both secondary waves being extraordinary. We remind that in uniaxial crystals the ordinary and extraordinary waves differ, first, in the shape of the refractive index surface and, second, in polarization: the extraordinary wave is polarized in the cardinal section plane while the ordinary wave has polarization normal to this plane (see Sec. 12).

In order to obtain secondary light waves at frequencies ν_1 and $\nu_2 = \nu - \nu_1$, it is necessary to orient the nonlinear crystal inside the resonator in such manner that the optical axis *OO* coincides with the synchronism direction for the chosen combination of frequencies ν and ν_1.

Methods of tuning parametric light oscillators. There are several methods of controlling frequency of secondary waves generated in parametric light oscillators.

Angle tuning. Let a nonlinear crystal be oriented inside the resonator in such way that its principal axis *AA* is at an angle α_1 to the optical axis *OO* (Fig. 73). In this case frequencies of the secondary waves, ν_1 and $\nu - \nu_1$, will be such that the synchronism direction will make the angle α_1 with the principal axis of the crystal. Let us slightly tilt the crystal, until the angle between the principal axis of the crystal and the oscillator axis

is α_2 (dashed line in Fig. 73). The frequencies of secondary waves that are generated now are such that the synchronism direction is at the angle α_2 to the principal axis of the crystal. It is thus possible to vary the frequency of secondary light waves (in a certain range, of course) by varying the angle α, that is by rotating the nonlinear crystal with respect to the axis *OO*.

Temperature tuning. We have mentioned in Sec. 12 that the refractive index of the medium is also a function of temperature. This means that in uniaxial crystals temperature must affect to some extent the shape of optical refractive index surfaces of the ordinary and extraordinary waves. This must lead, in turn, to changes in the direction of wave synchronism for any fixed combination of frequencies.

Consequently, the secondary wave frequency may be tuned by varying crystal temperature, without rotating the nonlinear crystal inside the resonator. As the crystal is heated or cooled, new synchronism directions, corresponding to new combinations of

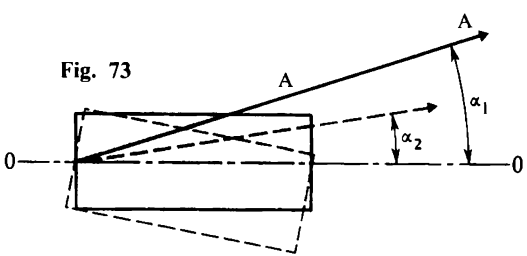


Fig. 73

secondary wave frequencies, will be coincident with the direction of the optical axis *OO*.

It should be noted in conclusion that both these methods of tuning are based, in the long run, on the dependence of the synchronism direction on the frequencies of interacting waves (dispersion of light is revealed in the dependence of the refractive index on frequency, see Sec. 12). The frequency of generated light waves is effectively controlled by directing the chosen direction of wave synchronism along the optical axis.

Is it possible to generate only one secondary light wave? A characteristic feature of the parametric light generation is the excitation of two secondary waves (at two distinct frequencies). Assume, however, that light of only one frequency must be emitted (say, ν_1). How to suppress the unwanted component of frequency $\nu - \nu_1$?

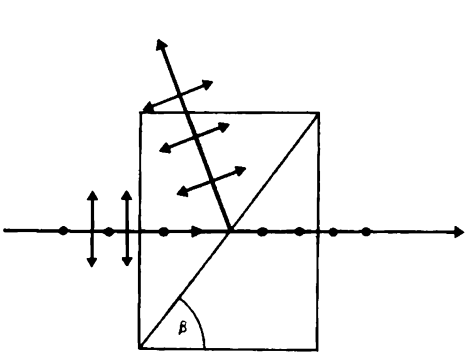
Obviously, the simplest approach would be to pass the exit beam through a filter which realizes resonant absorption at frequency $\nu - \nu_1$ but has almost no absorption at ν_1. Unfortunately, this would mean that part of the pumping wave energy consumed to excite the wave at frequency $\nu - \nu_1$, is completely lost; this energy will simply heat the filter.

A wiser approach consists in introducing into the resonator an additional element which substantially increases the level of losses for waves with frequency $\nu - \nu_1$ but does not appreciably affect that of the waves with frequencies ν and ν_1. This is realized with minimum difficulties if ν- and ν_1-waves are extraordinary and $\nu - \nu_1$ is an ordinary wave. In this case we can use the fact that polarization planes of the ordinary and extraordinary waves are mutually perpendicular.

It is convenient to employ the so-called Glan prism as the above-mentioned additional element. The Glan prism (Fig. 74) comprises two identical rectangular prisms cut out of an Iceland spar crystal in such fashion that the principal axis of the crystal is parallel to the edge of the angle β (i.e. normal to the plane of the drawing); $\beta = 50°$ The Glan prism has a very interesting property: it lets through, and without changing the direction of propagation, only the plane-polarized light wave with the plane of polarization normal to the plane of the drawing, while the light wave polarized in this plane is reflected away. The reflection takes place on the interface of the rectangular prisms.

Let us place a nonlinear crystal and a Glan prism along the optical axis *OO* inside the oscillator cavity as shown in Fig. 75 (*AA* is the principal axis of the nonlinear crystal, and the hatched plane is that of the cardinal section). The waves with frequencies ν and ν_1

Fig. 74

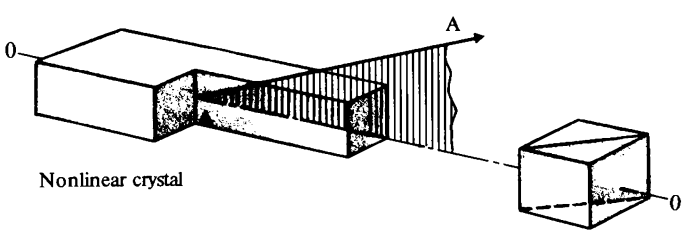


are extraordinary, so that their polarization plane coincides with the cardinal section plane. These waves will therefore be transmitted by the Glan prism. The $(\nu - \nu_1)$-wave is, on the contrary, an ordinary wave; it is polarized normal to the cardinal section plane and so will be eliminated by the prism.

A Glan prism in the oscillator cavity can therefore substantially affect the process of parametric generation of secondary light waves: the $(\nu - \nu_1)$-wave will not be generated at all and the pumping wave energy will be converted only to the wave with frequency ν_1. One important fact is that the Glan prism continues to function even if frequency is varied since the character of polarization of the interacting waves is obviously unaffected by the tuning operations.

Parametric light oscillator and optically pumped laser. The parametric light oscillator and optically pumped laser seem to be very similar instruments, at least in a number of characteristics. Both generate coherent light, and both employ an optical resonator

Fig. 75



(cavity). The similarity becomes more obvious if we take into account that a laser may be pumped by a coherent (laser) source and that the frequency of laser emission can be varied continuously (for example, in lasers with liquid lasants). If a laser is pumped by an auxiliary laser, then the functional diagrams of the two sources (the laser and the parametric oscillator) are nearly identical.

Despite the superficial similarity outlined above, the two sources differ in one essential point. The light-to-light transformation in the active medium of the laser proceeds via incoherent processes: the pumping wave is first absorbed by the matter and excites the active centres, and only then do the active centres emit light, thereby generating the secondary light wave. As to the parametric oscillator, its "active medium" is a scene of coherent processes: the pumping wave is "transformed" by a nonlinear crystal into secondary light waves, and no changes can be detected in the state of the crystal. The difference between the optically pumped laser and the parametric oscillator lies therefore in the difference between incoherent and coherent processes of light-to-light transformation.

The same factors determine the difference between "active mediums" in the sources in question. The parametric oscillator uses a nonlinear crystal specially oriented with respect to the resonator's optical axis; this "active medium" contains no active centres. On the contrary, lasers need no nonlinear mediums but the presence of active centres in them is essential (the system of energy levels, concentration of centres, the nature of their interaction with other particles of the lasant, etc.). In the case of the parametric oscillator the medium as a whole participates in the process, while in

the laser the "participants" are active centres, with specific quantum transitions occurring in each.

In contrast to the parametric oscillator, coherence of the pumping wave of a laser is not a necessary condition. Normally optical pumping involves incoherent light. Furthermore, pumping is in many cases not optical at all. For instance, excitation of active centres of gas lasers is often realized by means of inelastic collisions of particles in the plasma of gas discharge. Obviously, in this case even a superficial similarity of lasers and parametric oscillators vanishes.

19. Nonlinear Optics and Progress in Laser Technology

The best debt is the one paid back. Nonlinear optics and laser technology are closely interdependent fields. On one hand, nonlinear optics could develop only on the basis of laser successes. On the other hand, it is nonlinear optics that at present determines to a great extent further potentials of laser technology. "Born of the laser", nonlinear optics now opens new horizons to the laser applications. Indeed, the best debt is the one paid back.

Phototropic gate discussed in Sec. 14 gives an example of the effect of nonlinear optics on the laser field. Several years ago giant light pulses were obtained by various optical shutters of revolving type. Various "vanes" could be found in the laboratories — rotating beam choppers, rotating resonator mirrors or prisms,

and so on. Now such gates are replaced almost completely by more efficient devices, and among them by phototropic gates employing the effect of induced transparency of the medium. Advantages of such devices (and first of all their automatic operation) were described in Sec. 14.

Nonlinear optics serves to develop new types of sources of coherent light. As we mentioned in Section 5, an important problem is the expansion of the range already "mastered" by coherent-light generators. One of the most promising methods of solving this problem is found in making use of frequency converters of nonlinear optics. Among them are

generators of optical harmonics, and first of all the second-harmonic generators (frequency doublers);

parametric light oscillators;

oscillators for various operations of frequency mixing (summation, subtraction, combining, and so on).

All these devices are based on nonlinear optical phenomena, that is on various coherent processes of light-to-light transformation.

As an example, let us consider two diagrams of practically employed methods of successive frequency conversion. The *first* of them is shown in Fig. 76. A laser emits a coherent light wave with frequency ν. This wave is used to pump the parametric oscillator which emits radiation that can be continuously tuned in some frequency range, from $\nu_1 - \Delta\nu$ to $\nu_1 + \Delta\nu$. The output of the parametric oscillator serves as the pumping wave for the second-harmonic oscillator; the final result is a coherent radiation with frequency continuously tunable from $2\nu_1 - 2\Delta\nu$ to $2\nu_1 + 2\Delta\nu$.

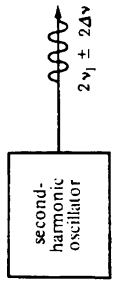

ν
parametric oscillator
ν_l ± Δν
second-harmonic oscillator
2ν_l ± 2Δν

Fig. 76

The *second* diagram is shown in Fig. 77. Laser emits a wave with frequency ν. After passing the wave through two second-harmonic oscillators we obtain a wave with frequency 4ν, that is the fourth optical harmonic. Mixing this harmonic with the initial signal (i.e. summing up $4\nu + \nu = 5\nu$) we obtain the fifth optical harmonic.

It must not be overlooked, however, that conversion at each stage of the system is far from total. Energy is inevitably lost at each of them, and these losses determine practical value of each specific configuration of the system. Normally a system is considered acceptable if from 10 to 20% of the initial-wave energy is converted to the secondary wave at each stage. In some cases the conversion coefficient was substantially improved, up to 0.3-0.5 (and even higher).

The feature common for schematics of Figs. 76 and 77 is that all instruments function independently and start operating on a sequential basis: first one, then the other, then the third. The laser as the source of the initial pumping wave is at the beginning of this "chain"

Although this principle of organization of the conversion appears to be quite logical, it is by no means optimal. If, for instance, the chain involves three successive stages of conversion, then even for the conversion coefficient at each stage equal to 0.5 the ultimate conversion coefficient will only be $(0.5)^3 = 0.125$. The following question is thus in order: is it possible to drop the principle of sequential (successive) frequency conversion? Is there an essentially

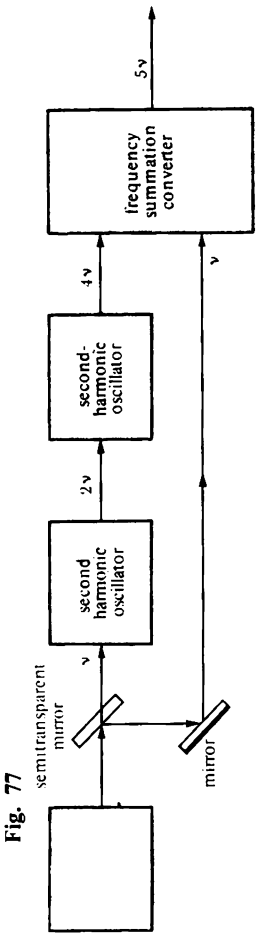

5ν
frequency summation converter
4ν
second-harmonic oscillator
2ν
second harmonic oscillator
ν
semitransparent mirror
mirror
ν

Fig. 77

new approach substantially increasing the efficiency of nonlinear frequency converters?

Fruitful idea. An essentially different approach does exist. In fact, it has been described in this book in at least three places. We want to remind about all three, as it is justified by the importance of the problem.

Case 1. Utilization of a prism inside the laser cavity was analyzed in Sec. 4. In order to suppress the unwanted wave with frequency ν_2, it could be possible to place the prism in the path of the beam propagating outside the cavity, and thus spatially separate the waves with frequencies ν_1 and ν_2. This would mean, however, wasting the energy consumed for generation of an unwanted wave. It is expedient therefore to place the prism inside the cavity and thus influence the process of light generation, that is make the appearance of the wave with frequency ν_2 impossible.

Case 2. It was explained in Sec. 4 why the end faces of the laser tube are oriented at the Brewster angle. It could be possible, in order to suppress the unwanted wave polarized normal to the "generation plane", to place a plane-parallel plate oriented at the Brewster angle across the beam emitted from the resonator. The plate would let through only the wave polarized in the "generation plane", but the energy spent on generation of the unwanted component would be lost. Therefore it is expedient to introduce the plate tilted at the Brewster angle into the cavity and at the same time make it the outlet window of the gas-discharge tube. This intrusion into the process of

generation damps out the excitation of a wave with polarization normal to the "generation plane"

Case 3. The possibility of generating only one secondary light wave was discussed in Sec. 18. An unwanted component with frequency $\nu - \nu_1$ could be removed by placing a Glan prism in the path of the light beam leaving the resonator. But again this would mean the loss of energy spent on generation of the unwanted component. This is why the Glan prism is placed inside the cavity which means intrusion into the process of parametric excitation and suppression of the generation of a wave with frequency $\nu - \nu_1$.

All the three examples provide a conclusive demonstration of the basic idea: it is much better to influence the process of generation inside the resonator than to operate with the beam already out of the resonator.

What is the "internal generation of the second harmonic"? The above idea may prove very fruitful for frequency conversion of coherent light. This means that harmonic generation). Then this wave is used to pump harmonic generation, should be placed inside the resonator.

Let us have a look at Fig. 78. The figure compares two qualitatively different situations: *a* — second harmonic generation with laser pumping, and *b* — *internal* SHG.

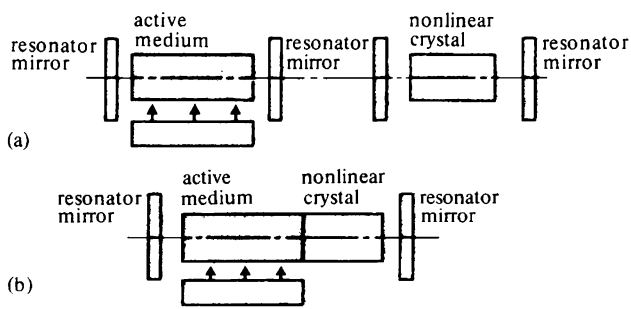
In the case of Fig. 78*a*, first a coherent light wave at frequency ν is obtained from the laser (first the nonlinear crystal meant, for instance, for the second the second harmonic oscillator, whereby part of the pumping energy is converted into a wave at frequency

2ν, that is into the second harmonic. If we denote the conversion coefficient by k, then the second harmonic intensity will be kI where I stands for the first harmonic intensity.

Figure 78*b* shows the case of internal generation of the second harmonic. In this configuration the wave with frequency ν may be totally absent. Photons with energy $h\nu$ are almost immediately involved into the three-photon coherent processes giving rise to emission of secondary photons with energy $2h\nu$. Studies demonstrate that the intensity of the second harmonic emission may reach the level of the intensity I of the first harmonic which would be emitted if the nonlinear crystal were removed from the laser resonator.

The next step is possible.... One could go even further. Why stop at placing a lasant and a nonlinear crystal in the same resonator and not to try and

Fig. 78



combine the two mediums into one? We mean a nonlinear uniaxial crystal with specially added active centres. Such a step may prove a logical extension of the previous development. It is too early yet to assess the consequencies of such a step. Still, activated nonlinear crystals are being grown already.

We find therefore that nonlinear optics greatly influences the laser field and opens ways of developing new types of lasers. At the same time, the "laser frequency range" is widened, the intensity of coherent emission enhanced, and its coherent properties improved.

Historical Background

We are near the end of the story about the laser age in optics. To bring this story to a close, let us scan the history of the field: some more important dates and names, the steps leading to the development of the laser, optical holography, and nonlinear optics.

The history of the laser. The word "laser" is an acronym of the words Light Amplification by Stimulated Emission of Radiation. The very term reflects, therefore, the fundamental role of the stimulated emission in generators and amplifiers of coherent light. Our outline of the history must logically start with 1917 when *A. Einstein* introduced the concept of stimulated emission.

This was the first step on the way to the laser. The next step was made by *V Fabrikant* in the USSR who indicated in 1939 that stimulated emission could be used for amplification of the electromagnetic radiation passing through the medium. Fabrikant's idea consisted in using microscopic systems with inverted

population. After the end of World War II, Fabrikant returned to this idea and in 1951 applied (together with *M. Vudynsky* and *F. Butayeva*) for the inventor's certificate of a method of amplification of electromagnetic radiation by making use of the stimulated emission. The subject of invention as formulated in the certificate was: "A method of amplification of electromagnetic radiation (in the ultraviolet, visual, infrared, and radio frequency ranges) having as its distinct feature the transmission of the radiation to be amplified through the medium in which an excessive, compared to the equilibrium, concentration of atoms or other particles and their systems at upper energy levels corresponding to the excited states of the said medium is produced by an additional radiation or by other means"

Originally, this method of amplification was realized in the radio frequency range (to be precise, in the microwave range). The Soviet physicists *N. Basov* and *A. Prokhorov* presented a report to the USSR Conference on Microwave Spectroscopy about the possibility of developing, in principle, a microwave amplifier. Basov and Prokhorov suggested the term "molecular generator" (a beam of ammonia molecules was to be employed). The suggestion to use stimulated emission to amplify and generate microwave radiation came at practically the same time from the American physicist *Ch. Townes* of the Columbia University.
A molecular generator later termed *maser* (Microwave Amplification by Stimulated Emission of Radiation) was realized in 1954. It was developed simultaneously and independently in two points of the globe: in the Lebedev Physics Institute of the USSR Academy of Sciences (by N. Basov and A. Prokhorov and their

co-workers) and in the Columbia University in the USA (by Ch. Townes and his co-workers).

The term "laser" is a later derivation from "maser" [substitution of M (for Microwave) in the acronym by L (for Light)]. Both the maser and the laser are based on the same principle formulated by Fabrikant in 1951.

The advent of the maser manifested that a new direction is gaining strength in today's science and technology. The field was first called *quantum radiophysics*, and then *quantum electronics.*

Ten years later, in 1964, A. Prokhorov said on the ceremony of presentation of Nobel Prize that everybody anticipated the appearance of the quantum light generators soon after masers had been developed. The prediction proved wrong, it took five to six years to develop the first laser. The reason for this lag, said Prokhorov, was two-fold: first, resonators (cavities) for the optical frequency range were not yet proposed, and second, there were no realistic systems and techniques in the optical range for realizing inverted population.

The five-to-six year span mentioned by Prokhorov was indeed consumed by the research which in the long run made it possible to bridge the gap between the maser and the laser. In 1955 Basov and Prokhorov substantiated the application of optical pumping to achieve population inversion. In 1957 Basov suggested that semiconductors be used to develop quantum light sources; he suggested, moreover, that specially machined surfaces of the semiconductor sample itself be used as the resonator walls. The same year Fabrikant and Butayeva recorded quantum amplification in the optical range in experiments with electric discharge in a mixture of mercury vapor with small amounts of hydrogen or helium. The next year Prokhorov, and

quite independently *A. Shavlov* and *Ch. Townes*, gave a theoretical basis for the utilization of the stimulated emission in the optical range. In 1959 *N. Basov*, *B. Vul*, and *Yu. Popov* published a paper theoretically substantiating the idea of developing semiconductor quantum light sources (semiconductor lasers) and analyzing a number of necessary conditions. Finally, N. Basov, *O. Krokhin*, and Yu. Popov published a lengthy paper in 1960 in which they discussed in detail the principles of developing quantum generators and amplifiers in the IR and visible ranges. The authors concluded: "There being no principal restrictions, one can hope that generators and amplifiers in the IR and visible ranges will be developed in the nearest future"

The intensive theoretical and experimental research effort in the USSR and USA has brought the scientists to the very "brink" of the laser development by the end of the 50's. The American physicist *T Maiman* was the first to report the success. In August and September 1960 he published communications in two British journals on the achieved emission in the visible range. The world was thus informed about the birth of the first "optical maser", the ruby laser. The first laser looked quite modest: a small ruby cube ($1 \times 1 \times 1$ cm in dimensions) with two opposite faces coated with silver film (acting as cavity mirrors). The crystal was periodically illuminated with green light of a high-power flash lamp which was wound, like a snake, into a helix around the ruby cube. Red light pulses of the emitted radiation were let out through a small hole in the silver coating of one of the faces.

The same year the American physicists *A. Javan*, *W. Bennett*, and *D. Herriot* succeeded in achieving

generation of optical radiation from the gas discharge (the active medium was a mixture of helium and neon gases). This was the first gas laser whose appearance was in fact prepared by the 1957 experimental results of Fabrikant and Butayeva.

Beginning with 1961, the laser (a solid state or a gas laser) becomes a permanent piece of equipment in optics laboratories. New lasants are being developed and the laser design technology is improving. First semiconductor lasers appear simultaneously in the USSR and in the USA in 1962-63.

This was the beginning of the "laser age" in optics.

To the history of optical holography. It is significant that the main ideas, principles, and techniques of holography were suggested long before the laser was developed. The advent of the laser in the 60's immediately provided a solid foundation to the holographic ideas and techniques; a new direction in optics, *optical holography*, became a reality.

In principle, the idea of the holographic method of image formation was suggested and verified experimentally in 1920 by the Polish physicist *M. Wolfke*. In his paper, "On the Possibility of Obtaining an Optical Image of a Molecular Lattice" he demonstrated that X-ray diffraction patterns of a crystal can be used to reconstruct an optical image of this crystal. Unfortunately, Wolfke's publication did not find any response from his contemporaries in physics. Wolfke remains a forerunner of holography who was neither understood nor supported while he was active.

The ideas and principles of holography were formulated anew in 1948 by the British physicist

D. Gabor (future Nobel Prize winner for his contribution) who at the time was not aware of Wolfke's results. Gabor's contribution to holography is exceptionally large; the term *holography* was also his suggestion. Incidentally, Gabor came to holography while working at a quite practical problem, namely, he was developing methods of improving resolution of the electron microscope. Gabor mentioned that the idea of the holographic method in electron microscopy as a two-stage process, in which an object is recorded by an electron beam and is reconstructed by a light beam, has appeared as a modification of an idea of the well-known British physicist *W. Bragg*. In his paper "X-Ray Microscope" Bragg presented in 1942 a method of visualizing the crystal lattice by means of diffraction of light on a diffraction pattern recorded with X-rays.

With the invention of the laser in 1960, the principles of holography were soon used to develop a serious and promising field of science. In 1961 the American physicists *E. Leith* and *J. Upatnieks* developed the widely used method of the two-beam holography in which the reference laser beam is used together with the object beam.

In 1962 the Soviet physicist *Yu. Denisyuk* suggested the idea of the three-dimensional holograms based on thick photoemulsion layers. He also developed a method of recording these holograms by using the back scattered object beam. Denisyuk's three-dimensional holograms reconstructed in the ordinary sun light are a result of many years of research started long before the invention of the laser as well as before the fundamental work of Gabor. Denisyuk mentions that for him the starting point was the method of colour photography developed as early as 1892 by the French

physicist *G. Lippmann.* Later, in 1970, the sequence of Denisyuk's papers under the general heading "Holography with Recording in a Three-dimensional Medium" was conferred the Lenin Prize for Science.

Almost simultaneously a number of laboratories in the USSR, USA, and England began to study holographic interferometry and immediately demonstrated its high practical significance and great potentialities. Holography was maturing.

We want to emphasize that although the holographic principles were formulated long before the laser age, the optical holography grew into a serious scientific and technological field only owing to progress in the laser science. As Denisyuk notes, "Holography without lasers would remain an interesting principle which possibly could be applied to some special problems. The laser gave holography a new life, opened for it numerous inroads into practical applications"

To the history of nonlinear optics. Two Soviet physicists, *S. Vavilov* and *V Levshin*, conducted in 1925 a very interesting experiment. They observed in experiments with high-intensity light (high-density spark discharge was used as the light source) a decrease in the uranium glass absorption coefficient by 1.5%, with the average experimental error of $\pm$ 0.3%. In fact, it was the first ever observation of a nonlinear effect in optics. It was the observation of the light-induced transparency of the medium.

At that time, however, these experiments had no future. The appropriate equipment, and first of all high-intensity light sources, was lacking. The pre-laser optics simply had nothing of the sort.

Despite this obstacle, Vavilov spent much time pondering the possibilities of analyzing nonlinear phenomena in optics. His exceptional scientific intuition, his ability to see far ahead into the future were demonstrated to the full in the book "Microstructure of Light" published in 1950. The main problems of the new field of optics and the approaches to solving these problems were outlined in this monograph with surprising insight and completeness. The name given by Vavilov to the new field was *nonlinear optics.*

We shall not be very wrong to say that originally nonlinear optics was born "at the tip of a pen", or in the scientist's mind. Vavilov did not live to see the laser-age progress in nonlinear optics; he died in 1951, ten years before the laser was invented.

The advent of the laser resulted in the rebirth of nonlinear optics. *P. Franken,* the American physicist, observed in 1961 generation of the second harmonic of the ruby laser radiation in a quartz crystal. An analysis of his results enabled two Soviet physicists, *R. Khokhlov* and *S. Akhmanov,* to formulate in 1962 the conditions of maximum efficiency for various nonlinear phenomena in optics (including generation of optical harmonics); they have advanced the idea of the parametric generation of light. Simultaneously, the American physicists *G. Giordmaine* and *R. Terhune* analyzed the possibility of satisfying the wave synchronism conditions in uniaxial crystals.

In 1961 the Soviet physicist *G. Askaryan* predicted the self-focusing effect; this prediction was confirmed by the experiments of *N. Pilipetsky* and *S. Rustamov* in the USSR who were the first to observe self-focusing of light (in 1963). In 1964 the American physicist

M. Hercher was the first to report the superthin "filaments" in self-focused beams.

Fundamental theoretical work on nonlinear optics was carried out in the period from 1961 to 1963 in the USSR by R. Khokhlov and others, and in the USA by *N. Bloembergen* and his co-workers. In 1965 Akhmanov and Khokhlov have published a fundamental monograph "Problems in Nonlinear Optics"

By 1965 nonlinear optics was established as an independent powerful branch of the contemporary optics. Sufficiently effective generators of optical harmonics and tunable parametric light oscillators were already available.

TO THE READER

Mir Publishers welcome your comments on the content, translation and design of this book.

We would also be pleased to receive any proposals you care to make about our future publications.

Our address is:

Mir Publishers
2 Pervy Rizhsky Pereulok,
I-110, GSP, Moscow, 129820
USSR

Printed in the Union of Soviet Socialist Republics

Basic Concepts of Quantum Mechanics

by L. TARASOV, Cand. Sc.

This book gives a detailed and systematic exposition of the fundamentals of non-relativistic quantum mechanics for those who are not acquainted with the subject. The character of the physics of micro-objects and the problems of the physics of microprocesses (interference of amplitudes, the principle of superposition, the specific nature of measuring processes, casuality in quantum mechanics) are considered on the basis of concepts about probability amplitudes. Besides, the quantum mechanical systems — micro-objects with two basic states are analyzed in detail. The apparatus of quantum mechanics is considered as a synthesis of concepts about physics and the theory of linear operators. A number of specially worked out problems and examples have been included in order to demonstrate the working of the apparatus.

This book is meant for use by students of engineering and teachers-training institutes. It may also be used by engineers of different profiles.

L. Tarasov's book "Laser Age in Optics" is a popular science book concerned with new trends in optics arising from development of the laser.
L. Tarasov, Cand Sc, who is an Assistant Professor of the chair of Electronics of the Moscow Institute of Electronic Machine Design, also teaches at the Moscow state University.
In this book the author successfully combines a popular style of presentation with, when necessary, a serious scientific treatment of the problems involved. Consequently, the book can be recommended not only for secondary school students but also for a layman adult reader wanting to study in some detail the principles involved in lasers, optical holography and nonlinear optics The reader will find a popular, vivid and intelligent presentation of many elusive concepts which, as a rule, are either not covered at all in literature of this kind or only in passing.